The Pennsylvania State University
The Graduate School

MITIGATING SOCIAL CHALLENGES AMONG VULNERABLE COMMUNITIES WITH MACHINE LEARNING

A Dissertation in
College of Information Sciences and Technology
by
Maryam Tabar

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2023

The dissertation of Maryam Tabar was reviewed and approved by the following:

Amulya Yadav
PNC Technologies Career Development Assistant Professor in the College of Information Sciences and Technology
Dissertation Co-Advisor, Co-Chair of Committee

Dongwon Lee
Professor in the College of Information Sciences and Technology
Dissertation Co-Advisor, Co-Chair of Committee

C. Lee Giles
David Reese Professor in the College of Information Sciences and Technology

Carleen Maitland
Professor in the College of Information Sciences and Technology

S. Shyam Sundar
James P. Jimirro Professor in the Donald P. Bellisario College of Communications

Rayid Ghani
Professor in Machine Learning Department at Carnegie Mellon University
Special Member

Jeffrey Bardzell
Professor in the College of Information Sciences and Technology
Associate Dean of Undergraduate and Graduate Studies

Abstract

There are various environmental and social challenges that disproportionately affect vulnerable communities in society. Extensive research has been conducted in various fields, such as agricultural sciences and social sciences, to understand some of those challenges and design intervention/prevention programs. However, effective/efficient implementation of mitigation plans is usually highly challenging in the field. Inspired by recent advances in Machine Learning (ML), this dissertation mainly focuses on the adaptation of ML-based techniques in certain real-world domains under various challenges to help address several social problems in a more effective/efficient manner. In fact, it focuses on two real-world domains, AI for Agriculture and AI for Social Welfare of Housing-Insecure Low-Income Americans, and addresses some challenges by proposing solutions tailored to the characteristics of the motivating problem domain. For example, to address the challenge of a lack of ground-truth labels, it proposes a label generation approach that translates the findings of social science research to high-quality labels to facilitate training ML models. Additionally, it proposes a loss function to improve the learning of neural networks when only coarse-grained ground-truth labels are available. In conclusion, this dissertation aims to adapt ML algorithms in specific real-world domains with particular challenges and characteristics.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This material is based upon work in part supported by the Bill and Melinda Gates Foundation under Award No. #141840 and the NSF under Award No. #1742702, #1820609, #1915801, #1934782, #1909702, #1940076, and the High-Potential Individuals Global Training Program (2019-0-01590) by IITP and MSIT, Korea. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

Chapter 1
Introduction

In 2015, United Nations designed 17 Sustainable Development Goals (SDGs)[1], which are a call for action to build a better future by addressing various global challenges. In these SDGs, particular attention is paid to the needs of/challenges faced by vulnerable communities to improve their lives from multiple perspectives. For example, SDG#1 ("No Poverty")[1] calls for action to end poverty and ensure that all vulnerable populations have access to basic services, etc. Inspired by this global movement, this dissertation proposes Machine Learning (ML) based solutions to help mitigate several social challenges faced by certain vulnerable populations more effectively and efficiently; in particular, it focuses on two vulnerable populations, namely, smallholder farmers in Sub-Saharan Africa and housing-insecure low-income Americans, and aims to help mitigate some of the challenges that they struggle with (Figure 1.1 represents the real-world problem domains motivated my Ph.D. work).

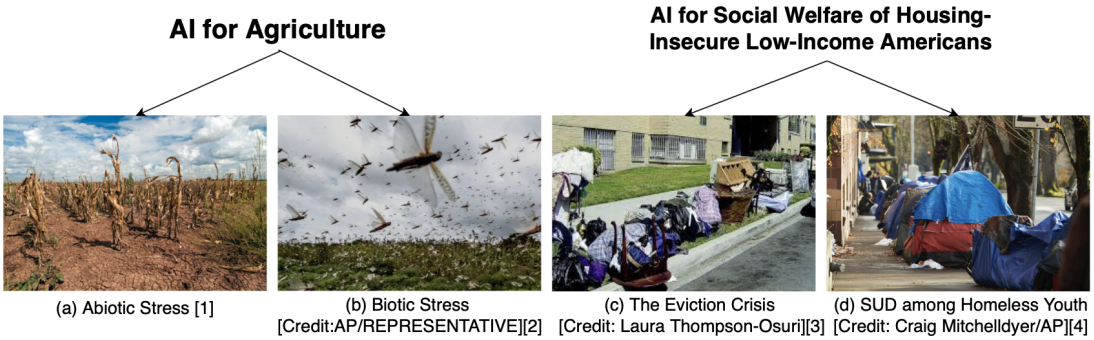


Figure 1.1: Real-world domains that have motivated my Ph.D. research.

[1] `https://sdgs.un.org/goals`

From a technical perspective, this dissertation mainly focuses on the adaptation of ML techniques in real-world domains with particular challenges that mainly stem from the characteristics of their data, such as the availability of coarse-grained ground-truth labels alone (rather than fine-grained/individual-level labels) [5], the absence of any ground-truth labels [6], etc. In fact, depending on the nature of the social challenge of study and target variable, this dissertation incorporates various data sources corresponding to the characteristics of individuals and/or environments, which in turn, are usually collected through surveying[2] and remote-sensing (such as satellite images), respectively. However, sometimes, the available data has certain characteristics and/or limitations that could adversely affect the performance of some existing ML models, and hence, need particular attention. For example, sometimes, the relevant input factors are not necessarily available at the spatial/temporal resolutions of interest because they might be collected by various agencies for specific purposes, and also, collecting such data might not be possible for each data scientist. Furthermore, sometimes, due to the data collection methodology (such as crowdsourcing human volunteers) or the existing obstacles to data acquisition, the input data might be significantly sparse, or the target variable might not be available at the resolution of interest. Therefore, it is important to consider the limitations/challenges that the data introduces to the application of ML models in the real world. Accordingly, this dissertation works on some of such real-world challenges, and studies the performance/weaknesses of several existing ML models on real-world data with particular characteristics. Then, to address existing challenges, it proposes solutions tailored to the constraints/characteristics of the motivating problem domains. In the following sections, we describe the studied social challenges, characteristics/limitations of the available data, and our solutions briefly.

[2]In our work, depending on the granularity of the target variable (individual-level or neighborhood-level), the granularity of input survey data might change. For example, Chapters 4-6 rely on neighborhood-level survey data as their proposed ML models predict the eviction condition at the neighborhood level, while Chapter 7 relies on the individual-level survey data because it works at the individual level.

1.1 AI for Agriculture

Smallholder farmers and their farms form the backbone of agriculture and food security in Africa. Unfortunately, due to the prevalence of some biotic/abiotic stresses, they deal with low crop-productivity, which in turn, significantly affects their livelihoods and food security [7]; e.g., agriculture on smallholder farms is an important means of livelihood for over 60% of individuals in Sub-Saharan Africa [8–11]. Therefore, supporting this vulnerable population and their agriculture could play a key role in poverty eradication and increasing food security which are the focus of SDG#1 ("No Poverty")[1] and SDG#2 ("Zero Hunger")[1], respectively [12].

To help mitigate these issues, this dissertation proposes data-driven solutions to forecast the occurrence of some abiotic and biotic stresses in order to make better proactive plans. In particular, Chapter 2 focuses on predicting three important crop-productivity related variables (namely, net primary production, actual evapotranspiration, and reference evapotranspiration), which in turn, could help get to know the occurrence of some abiotic stresses (such as drought) ahead of time. Then, Chapter 3 focuses on forecasting the presence of Desert Locusts (as a highly destructive biotic stressor) from crowdsourced data as well as relevant environmental factors. From a technical perspective, Chapters 2 and 3 work on addressing two main real-world data challenges, namely Data Variability and Sparse/Non-Uniformly Distributed Data, in the aforementioned problem domains. In the following paragraphs, we briefly describe the studied challenges and our solutions.

Learning under Data Variability. High variability has been seen in time-series data of a wide region in various domains. In particular, high variability can be observed when studying crop productivity-related factors across a wide region [13] because those factors are usually highly associated with climate conditions, crop type, soil type, and other environmental characteristics of a region, which in turn, could change considerably across a wide region (such as a continent). Therefore, it is important to effectively model such variability to build an accurate predictive ML model in such domains. Accordingly, focusing on the agriculture domain, Chapter 2 develops a predictive algorithm that models variability in time-series data through a clustering-based approach. Further, it studies the capability of

Variational Recurrent Neural Networks (VRNNs) [14] in capturing variability, and as a result, it empirically finds that training VRNNs sometimes might lead to the Posterior Collapse issue [15–19], which can hinder their capability in capturing variability in practice, although they have such a capability in theory.

Learning from Sparse/Non-Uniformly Distributed Data. High data sparsity could pose a major challenge in the ML domain, and the data collection methodology and its characteristics play an important role in data sparsity. Accordingly, Chapter 3 focuses on a domain, in which time-series data is collected through crowdsourcing (rather than sensors) across a wide region, and hence, the data is not distributed uniformly across space/time (at many points in time, there is no data available from many regions). Then, to address this issue, it builds a feature representation approach that turns the crowdsourced data into a suitable form based on the characteristics of the motivating problem domain. Then, it also develops a model-agnostic data augmentation approach to further address the challenge of data sparsity in that specific domain.

1.2 AI for Social Welfare of Housing-Insecure Low-Income Americans

Many low-income individuals in the United States are at a high risk of eviction and/or homelessness, partly due to a lack of affordable housing and a gap between income growth and increases in housing expenses [20, 21]. These crises could adversely affect their lives from multiple perspectives such as health, education, employment growth, etc. [20, 22]. Therefore, addressing the needs of this vulnerable population is critically important, and in turn, could make strong contributions to advancing towards several SDGs such as SDG#1 ("No Poverty")[1] and SDG#11 ("Sustainable Cities and Communities")[1]. To this end, this dissertation proposes AI-powered solutions to help mitigate some challenges faced by this population, namely eviction and substance use disorder, more effectively/efficiently.

From a technical perspective, this dissertation, in part, proposes computational approaches to help improve the state of practice under various situations in terms of the level of access to ground-truth labels. In fact, the availability of large-scale labeled data plays a key role in the success of various ML-based systems in the real

world. However, such data is not necessarily available at the resolution of interest in the field, and collecting ground-truth labels at scale is sometimes highly costly and time-consuming in many real-world situations. Therefore, it is important to facilitate the process of learning ML models under a lack of ground-truth labels of interest. Accordingly, this dissertation, in part, focuses on situations, where no ground-truth label is available or their resolution is lower than the prediction resolution, and relies on an interdisciplinary approach to address these challenges. In the following paragraphs, I will briefly describe the studied challenges and our solutions.

Learning from Coarse-Grained Ground-Truth Labels. Sometimes, the granularity of available labels is coarser than the granularity of prediction; i.e., it is highly difficult to obtain the ground-truth label for each individual training data point, however, the labels of a group of data points (in the training set) are easily accessible. This challenge can hinder the straightforward application of fully-supervised approaches in those problem domains. Motivated by this challenge, Chapter 5 focuses on one of such real-world domains, where an accurate regression model is needed and the spatial resolution of labels is lower than the spatial resolution of prediction. It studies the capability of existing research [23], and finds them to not perform well in that specific problem domain. Then, it proposes a new loss function that (1) leverages low-resolution ground-truth labels to ensure that the model's prediction is accurate at a low spatial resolution, and (2) leverages high-resolution sociological insights to be able to differentiate various data points (i.e., capture variability among data points), and hence, have accurate predictions at a high spatial resolution as well.

Learning under Absence of Ground-Truth Labels. Sometimes, ground-truth labels are not readily available, and also, collecting labels at scale is highly costly and time-consuming. This, in turn, could pose significant challenges to the process of building ML models. Chapter 6 focuses on one of such real-world domains, where the desired output is a categorical variable, and it assumes that no ground-truth label is available during training. Then, it proposes a label generation approach that translates the findings of social science research to high-quality binary labels (in an unsupervised manner), which can then be used to train ML models of interest.

1.3 Overview of Dissertation

The remaining part of this dissertation is organized as follows. Chapter 2 focuses on modeling variability in time-series data and studies the capability of some variational neural networks in capturing variability in practice. Then, Chapter 3 develops feature representation and data augmentation solutions to facilitate the process of learning from sparse and non-uniformly distributed data in a specific domain. Additionally, Chapters 4, 5, and 6 propose various AI-driven solutions to help improve the state of practice in mitigating the eviction crisis under various levels of access to ground-truth labels. Finally, as a separate use case, Chapter 7 shows how ML could be used to help mitigate Substance Use Disorder among homeless youth.

Chapter 2
AI for Agriculture: Abiotic Stress / Crop Productivity Prediction under Data Variability

This chapter focuses on forecasting three important crop-productivity related variables with the goal of helping smallholder farmers in Sub-Saharan Africa get to know the occurrence of some abiotic stressors ahead of time, and hence, improve their productivity and profitability [11]. From a technical perspective, it works on the variability inside time-series data, and also, examines the capability of some variational neural networks [24] in capturing variability in practice. In the following sections, we describe the problem domain, related work, our solution, experimental results, and several real-world use cases that we envision for such a predictive algorithm.

2.1 Introduction

Smallholder farms (less than two hectares in size) and their farmers form the backbone of African agriculture and food security and constitute a significant proportion of the Gross Domestic Product (GDP) of several African countries. For example, agriculture on smallholder farms is the primary means of livelihood for more than 60% of people in Sub-Saharan Africa and is responsible for ∼75% of the region's total agricultural production [8–10]. In addition, smallholder agriculture also plays

a critical role towards meeting several Sustainable Development Goals (SDGs)[1] laid out by the United Nations, such as "no poverty and zero hunger". Thus, developing techniques to improve the productivity and profitability of these smallholder farms is of critical importance, as it could lead to significant improvements in the well-being of many disadvantaged communities in Africa.

Unfortunately, increasing the productivity/profitability of smallholder agriculture is a challenging problem because of several reasons: (1) smallholder farmers find it difficult to protect their farms against various stressors (e.g., pest and disease outbreaks); (2) they lack awareness about modern agricultural practices; and, most importantly, (3) over the last few decades, climate change on the African continent has significantly hampered the ability of smallholder farmers to achieve high agricultural productivity [25]. In fact, the high reliance of smallholder farmers on rain-fed agriculture, coupled with a lack of knowledge about future climatic conditions result in highly uncertain situations for farmers. For example, many farmers do not know the irrigation needs of their crops at any given point in time [26]. This is one of the primary factors behind consistently low agricultural productivity among the smallholder farmers. As such, it is of great importance to help them get a better understanding of future conditions on their farms, so that they can proactively assess and address their irrigation needs.

In this chapter, we tackle this important problem by developing CLIMATES (**Cl**ustering **I**nitialized **M**eta **A**lgorithm for **T**ackling **E**nvironmental **S**tressors), a predictive algorithm to forecast three important crop-productivity related variables: (1) actual evapotranspiration (AET); (2) reference evapotranspiration (RET); and (3) net primary production (NPP). Intuitively, AET and RET could be used to measure the amount of water present in soil/needed to support crop growth, whereas NPP could indicate the amount of crop growth that occurs inside a farm. Generating accurate predictions for these three variables can help smallholder farmers understand their irrigation needs better, e.g., if the AET forecast for a smallholder farm shows stress (i.e., the forecasted AET value is less than what is required for healthy crop growth), then a farmer can proactively start irrigating his/her farm to mitigate that stress.

In fact, in this chapter, we make the following main contributions: (1) In

[1] `https://www.un.org/sustainabledevelopment/`

collaboration with PlantVillage[2], we identify ∼2,200 smallholder farm locations across Africa, and gathered remote-sensed data for AET, RET, and NPP for all these farm locations; (2) We develop CLIMATES, which leverages cluster-based structural insights of environmental time-series data in this domain, and then uses a distinct predictive model to make (AET, RET, and NPP) forecasts for each cluster; (3) We conduct a comprehensive analysis of the effectiveness of various popular classical ML and deep learning methods for time-series forecasting and show that CLIMATES outperforms all these baseline models. In particular, we provide insights about why Variational Recurrent Neural Networks (VRNNs) [24], which explicitly model variability in sequential data, do not perform comparably.

This work is done in collaboration with PlantVillage. CLIMATES could serve as the engine of an early warning system for smallholder farmers who can use these warnings to proactively address the needs of their farms (such as irrigation needs).

2.2 Related Work

In this section, we discuss related studies in the agriculture and AI disciplines.

Agriculture Research. Numerous studies in the agriculture domain [27–29] have focused on estimating crop-productivity variables (i.e., AET, RET, and NPP) from meteorological factors, and finding associations between them. However, there have been a few attempts at using traditional models (such as KNN [30] and ARIMA [31]) and neural models [32] to predict ET. Typically, most of these studies focused on the data of a small region, which limits our understanding of their performance on the data of a wide region with large variability. In contrast, the focus of our work is to forecast NPP, AET, and RET in smallholder farms that span widely across Africa (and hence, large variability is expected due to various climate conditions, crop types, etc.). Furthermore, prior work found that statistical/ML models are more accurate than Historical Average methods (which do not involve learning). However, mixed results were reported when assessing the superiority of neural network models to other algorithms. For example, Izadifar [33] found that Multiple Linear Regression outperforms a neural network model in the task of predicting

[2]PlantVillage (`https://plantvillage.psu.edu`) is a non-proft working for assisting farmers in Africa in adapting to climate change and its consequences.

AET. However, this work relied on MLP as their neural network model, instead of using network architectures that were designed to explicitly model the sequential structure of time-series data, such as RNNs. In our work, we compare CLIMATES against stronger baselines such as VRNNs, Long Short-Term Memory (LSTM) [34], etc.

Artificial Intelligence Research. To the best of our knowledge, there have been no prior studies in the AI community on forecasting these three crop-productivity variables across a large geographic region. However, there has been a large body of research on modeling sequential data for time-series forecasting. Some models, such as SARIMA [35] and TBATS [36], focused on explicitly modeling certain statistical properties of time-series data. Some other work in the neural network domain focused on tackling various challenges in different time-series data; e.g., State Frequency Model (SFM) [37] combines the ideas behind LSTM and Discrete Fourier Transform to learn multiple frequency patterns in time-series data. In particular, one line of prior research focused on building deep latent variable models to capture variability in sequential datasets; for example, Jia et al. [13] and Chien and Kuo [14] proposed VRNN-based deep generative models for cropland detection and speech separation, respectively. In our work, we show that CLIMATES achieves higher predictive accuracy than many of these baselines.

2.3 Datasets

Through our collaboration with PlantVillage, we identified 2,264 smallholder farm locations across Africa (as shown with red dots in Figure 2.1). For each farm location, we collected remote-sensed time-series data for three variables (AET, RET, and NPP) over five years (from the beginning of 2015 to the end of 2019) from the WaPOR website[3], which is administered by the UN-FAO. For completeness, we provide a formal definition of these variables.

- **Actual Evapotranspiration (AET)**: AET refers to the summation of evaporation from soil, canopy transpiration, and interception. It can be used to derive the water demand of each crop; i.e., the difference between AET

[3] `https://wapor.apps.fao.org/home/WAPOR_2/1`

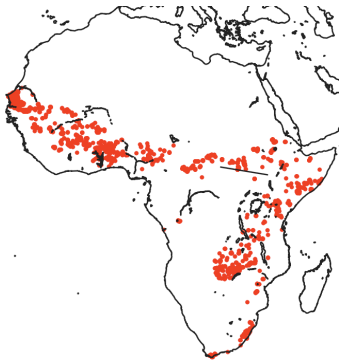


Figure 2.1: Identified farm locations across Africa

and RET (defined next) can be used to measure drought stress. Its unit is mm/day and its value ranges between 0.0 to 8.3 in our dataset [38].

- **Reference Evapotranspiration (RET)**: RET refers to the evapotranspiration of a well-watered plant under well-defined standard conditions. Its unit is mm/day and its value ranges between 1.1 to 12.7 in our dataset [38, 39].
- **Net Primary Production (NPP)**: NPP refers to the amount of carbon dioxide absorbed by plants, and is an indicator of plant growth. The unit of NPP is $gC/m^2/day$ (grams of carbon / square meter / day) and its value ranges between 0.0 to 9.265 in our dataset [38].

Data Characteristics. The WaPOR website provides data for AET, RET, and NPP, with a spatial resolution of 0.00223° (∼250 m) and a temporal resolution of one dekad (∼10 days) [38]. Using this data, we generate three separate time-series datasets (one for each AET, RET, and NPP). Each dataset consists of 2,264 time-series data points (each data point is the time-series for a specific farm location), and the length of each time-series is 180 (since we collect dekadal data over five years, i.e., $36 \times 5 = 180$). For each dataset, we consider the first three years of data (i.e., from the beginning of 2015 to the end of 2017) as the training set. The data in 2018 (and 2019) is kept as the validation (and test) set, respectively. As a pre-processing step, we apply Min-Max normalization on the data of each farm, however, predictive performance metrics are computed after converting the data back to its original scale.

2.4 The Meta-Algorithm: CLIMATES

In this section, first, we discuss key structural insights about our dataset which motivate the design of CLIMATES. Then, we describe our algorithm.

Exploratory Data Analysis. As shown in Figure 2.1, our 2,264 farm locations span widely across the African landmass. In total, these farm locations span across 20 different countries, each with its distinct climatic conditions. For example, while our farm locations in north-western Africa belong to the semi-arid Sahel region, farms in central Africa had tropical rainforest climates, and farms in eastern and south-eastern Africa had savannah grassland climates, etc.

Due to this geographic and climatic diversity across our farm locations, we expect significant *variability* in all three of our datasets. To investigate this further, we cluster each dataset (separately) using an off-the-shelf feature-based clustering approach [40]. At a high level, this clustering approach extracts the features of each time-series data point by applying Discrete Fourier Transform on its training portion. Once the feature vector for each time-series data point is extracted, bottom-up agglomerative clustering is used (with the Euclidean distance metric, and complete-linkage strategy for merging intermediate clusters).

As a result of data clustering, we obtain six clusters that have distinctly different shapes and patterns. Figure 2.2 illustrates three of these clusters obtained on the NPP dataset (we see similar results on the AET and RET datasets). Due to this significant variability, therefore, *we hypothesize that forecasting methods that may work well for data points in one cluster may not necessarily work well on other clusters*. This crucial insight motivates our design of CLIMATES.

The Proposed Meta-Algorithm. Given this strong variability inside our datasets, we conducted a cluster-by-cluster comparison of the predictive performance of several popular classical and deep learning-based forecasting methods. This analysis would help us understand whether a single forecasting method works best across all clusters, or whether different methods work better in different clusters.

For this comparison, we consider the time-series data points belonging to each of our six clusters separately. Then, on the data of each cluster, we train and test a heterogeneous mix of statistical, classical ML, and deep learning methods, namely TBATS, SARIMA, Linear Regression (LR), Random Forest (RF) [41],

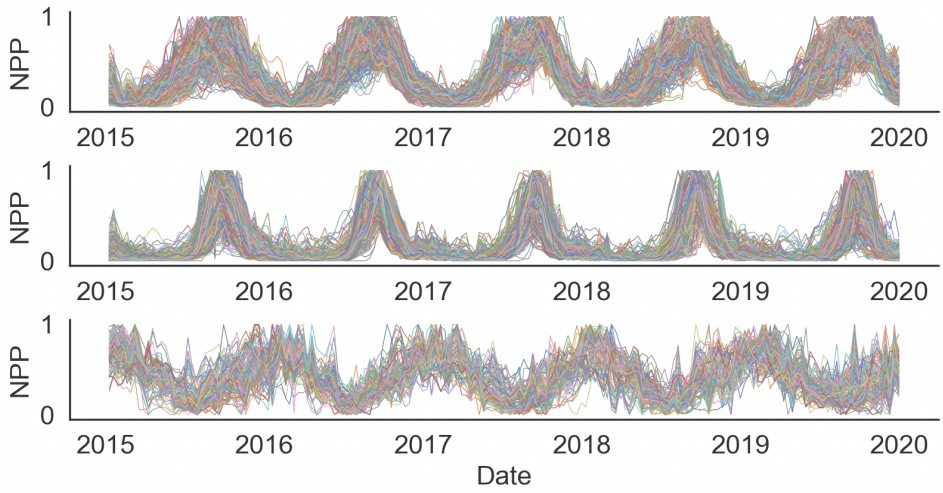


Figure 2.2: Three (out of six) NPP clusters generated through feature-based clustering

XGBoost [42], Support-Vector Machine (SVM) [43], LSTM, SFM, and Temporal Convolutional Network (TCN) [44].

Table 2.1 shows the coefficient of variation (CV)[4] achieved by the aforementioned methods on all six clusters found on the NPP dataset. (analogous results on the AET and RET datasets are represented in Tables 2.2 and 2.3, respectively). Note that these results are for single-step forecasting, i.e., we try to predict the next dekadal NPP, AET, and RET values. These tables confirm that no single forecasting method works best across all clusters, e.g., on the NPP dataset, statistical methods like SARIMA work best on the second and third clusters, deep learning methods like LSTM and SFM work best on the fourth, fifth, and sixth clusters, whereas Random Forest model works best on the first cluster. Thus, to get accurate forecasts consistently across the wide expanse of the African landmass, it is critically important to rely on an ensemble of well-trained models, each of which works well on a specific region of Africa.

Based on this finding, we now describe our meta-algorithm. CLIMATES works as follows: (1) It clusters the time-series data using a feature-based clustering approach into different clusters. (2) For each of these clustered datasets, it finds the best-performing forecasting model (i.e., the model with the lowest CV on

[4]Coefficient of variation (CV) refers to the root mean squared error divided by the average of the target variable. Therefore, the lower CV is, the better performance a method has.

Model	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
TBATS	0.2110	0.2860	0.2491	0.5235	0.4417	0.2279
SARIMA	0.1896	**0.2409**	**0.2112**	0.3840	0.3518	0.2071
LR	0.1731	0.2424	0.2234	0.3891	0.3603	0.2127
RF	**0.1726**	0.2481	0.2159	0.3822	0.3510	0.2123
XGBoost	0.1807	0.2555	0.2313	0.3992	0.3788	0.2249
SVM	0.1889	0.2582	0.2486	0.3916	0.4165	0.2301
LSTM	0.1728	0.2505	0.2349	**0.3774**	0.3446	**0.2035**
SFM	0.1740	0.2446	0.2199	0.3800	**0.3412**	0.2186
TCN	0.1890	0.2618	0.2410	0.3817	0.3774	0.2099

Table 2.1: CV of different models on the NPP clusters

Model	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
TBATS	0.1947	0.2285	0.2380	0.3060	0.3564	0.2217
SARIMA	**0.1713**	**0.2044**	0.2197	0.2799	0.2990	0.2049
LR	0.1763	0.2051	0.2179	0.2806	0.2981	0.1976
RF	0.1725	0.2058	0.2180	0.2772	0.2834	0.1976
XGBoost	0.1742	0.2104	0.2187	0.2839	0.2943	0.2012
SVM	0.1769	0.2162	0.2262	0.3044	0.3145	0.2002
LSTM	0.1723	0.2097	**0.2114**	**0.2669**	0.2728	0.1984
SFM	0.1767	0.2115	0.2115	0.2709	**0.2722**	**0.1967**
TCN	0.1764	0.2058	0.2263	0.2693	0.2883	0.2016

Table 2.2: CV of different models on the AET clusters.

the validation set of that cluster). We select the best-performing model on each cluster out of the nine models shown in Table 2.1. Note that we use this selection of models inside CLIMATES to ensure a good heterogeneous mix of statistical methods, classical ML methods, and deep learning methods. We further note that as more sophisticated time-series forecasting methods are developed, they can also be used as part of the CLIMATES ensemble. (3) At test time, each time-series data point is assigned to a subset of clusters. We considered two general strategies for assigning data points to the clusters: (a) we assign each time-series data point to the nearest cluster (CLIMATES-I), (b) we assign each time-series data point to a subset of clusters that falls within d distance from that data point. The threshold d is set to two heuristically computed values: (1) the average distance between the data points and their closest cluster (CLIMATES-II), (2) the median distance

Model	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
TBATS	0.1259	0.1435	0.1502	0.1029	0.0982	0.1663
SARIMA	0.1084	0.1261	**0.1297**	0.0910	0.0876	0.1552
LR	0.1014	0.1244	0.1509	0.0898	0.0885	0.1517
RF	0.0998	0.1233	0.1474	0.1050	0.0890	0.1496
XGBoost	0.1018	0.1229	0.1410	0.1038	0.0896	0.1515
SVM	0.1066	0.1307	0.1463	0.1148	0.0921	0.1508
LSTM	**0.0988**	**0.1180**	0.1372	**0.0839**	0.0820	**0.1430**
SFM	0.0991	0.1203	0.1346	0.0875	**0.0792**	0.1443
TCN	0.1022	0.1306	0.1365	0.0979	0.0850	0.1492

Table 2.3: CV of different models on the RET clusters.

between the data points and their closest cluster (CLIMATES-III). (4) Finally, the best-performing model on each selected cluster (in our chosen subset) is used to get a prediction on that test data point, and the average of the predicted values is returned as the final forecast of CLIMATES. We now conduct a rigorous evaluation of the predictive performance of our meta-algorithm against a comprehensive set of baselines.

2.5 Experimental Evaluation

We provide two sets of results. First, we provide a brief background on the VRNN architecture and show results comparing the predictive performance of CLIMATES against VRNNs, which at least in theory, should serve as a strong baseline. Second, we show results comparing the predictive performance of CLIMATES against a wide variety of statistical/classical ML and deep learning models.

VRNN Architecture. VRNN is a deep generative model that extends the idea behind Variational Autoencoders (VAE) [45, 46] to sequential data. VRNNs can be viewed as a sequence of VAEs conditioned on the hidden state of an RNN. Thus, similar to VAEs, they consist of generative and inference networks; the latter encodes the input into a latent space, and the former generates the output by reconstructing the input from the latent space. In fact, the generative process at time t begins with generating the latent variable z_t from a Gaussian distribution. However, unlike VAE, z_t is conditioned on h_{t-1} (the hidden state of RNN at time $t-1$) to be

Model	AET	RET	NPP
CLIMATES-I	**0.2075**	**0.0989**	**0.2409**
$\text{VRNN}^{KL}_{Deterministic}$	0.2161	0.1052	0.2594
$\text{VRNN}_{Deterministic}$	0.2166	0.1053	0.2639
$\text{VRNN}_{Gaussian}$	0.2836	0.1504	0.4496
$\text{LSTM}_{Deterministic}$	0.2113	0.1039	0.2617
$\text{LSTM}_{Gaussian}$	0.2754	0.1507	0.3863

Table 2.4: A CV comparison between CLIMATES and VRNNs/LSTMs

able to model the consistency within a single time-series data point [24]. During training, VRNN aims to maximize the log-likelihood of observations $\ell(p(x_{\leq T}))$, where $x_{\leq T} = \{x_1, ..., x_T\}$ represents the input time-series of length T. However, as inferring the log-likelihood is computationally intractable, VRNN maximizes the variational lower bound of the log-likelihood given in Equation 2.1. This lower bound consists of two terms: (1) reconstruction likelihood, and (2) the KL distance between the approximate posterior and the prior distributions. We compare CLIMATES against VRNNs because their ability to learn explicit representations of variability across time-series data points (through the sequence of latent variables $z_{\leq T}$) makes them ideal models for our domain.

$$\ell(p(x_{\leq T})) \geq \mathbb{E}_{q(z_{\leq T}|x_{\leq T})}[\sum_{t=1}^{T}(log(p(x_t|z_{\leq t}, x_{<t})) - KL(q(z_t|x_{\leq t}, z_{<t})||p(z_t|x_{<t}, z_{<t})))] \quad (2.1)$$

Comparison with VRNNs. We now provide results comparing the performance of CLIMATES against VRNN and LSTM. In this set of experiments, a separate LSTM and VRNN is trained (and tested) on each of our three datasets. For both VRNN and LSTM, we experiment with two different output functions (Deterministic and Gaussian). Finally, the negative of the variational lower bound given in Equation 2.1 is used as VRNN's loss function.

Table 2.4 compares the CV achieved by CLIMATES-I against VRNN and LSTM variants (for single-step forecasting on our three datasets). This table shows that regardless of the choice of output function, CLIMATES-I outperforms both VRNN and LSTM models. In fact, CLIMATES-I, on average, achieves 6.3% lower CV than

$\text{VRNN}_{Deterministic}$, even though VRNNs have latent variables to model variability inside our datasets. *Additionally, applying t-test, we find that the difference between the CV of CLIMATES-I and these models is statistically significant (p-value is consistently less than 0.01).* Surprisingly, these results show that VRNN is unable to outperform LSTM on any dataset; in particular, $\text{VRNN}_{Deterministic}$ (which has stochastic latent states) cannot outperform $\text{LSTM}_{Deterministic}$ (which does not have any stochastic components). Counterintuitively, this shows that explicitly learning representations of variability inside our datasets does not seem to help.

Why Do VRNNs Not Work? To understand VRNN's poor performance, we take a closer look at $\text{VRNN}_{Deterministic}$'s learning curves during training (we see similar results with other output functions). In particular, we separately analyze the learning curves of two components in the VRNN's loss function, i.e., (1) the reconstruction loss, and (2) the KL term. Figure 2.3 illustrates the changes in the values of these two components with the increasing number of epochs on the NPP dataset. According to this figure, the KL distance vanishes into zero after a few epochs; i.e., the approximate posterior becomes equal to the prior in the early epochs, and hence, the model starts ignoring latent variables in the early stages of training (we see similar results on other datasets). Thus, we observe that, in practice, training VRNN leads to a local optimum which hinders capturing variability in a dataset, even though, in theory, it has the capability of capturing variations. Similar findings have been reported with VAEs, e.g., prior research found that the same issue (called *"posterior collapse"*) occurs in VAE [15–19]. However, to the best of our knowledge, our work is the first to report this posterior collapse issue with VRNNs. Further, prior work proposed KL-annealing to tackle posterior collapse in VAEs [16]; however, as the second row of Table 2.4 (i.e., $\text{VRNN}^{KL}_{Deterministic}$) shows, even with KL-annealing, VRNN is unable to beat CLIMATES-I.

Comparison with Other Baselines. Having established the superior predictive performance of CLIMATES over VRNNs and LSTMs in Table 2.4, we now evaluate CLIMATES against the same baseline forecasting models that we used in Table 2.1, as all models there form the individual building blocks inside our CLIMATES approach. Note that we choose these algorithms as baselines in order to establish the effectiveness of our clustering based meta-algorithm approach over individual baselines. Further, we note that as more sophisticated time-series forecasting

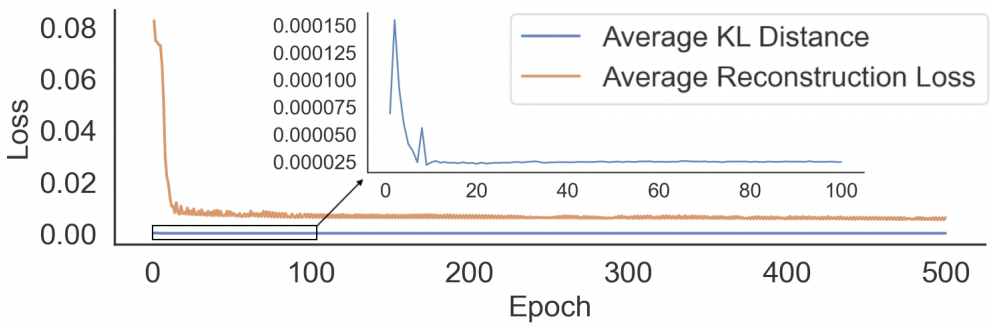


Figure 2.3: The learning curves of the two components of VRNN's loss function during training on the NPP dataset

methods are developed in the AI community, they can also be utilized as building blocks inside our meta-algorithm approach. To have a fair comparison between various models, we conduct hyper-parameter tuning using the grid search approach.

Table 2.5 shows the CV achieved by CLIMATES and all our baselines on single-step forecasting tasks on all three datasets. According to these results, CLIMATES outperforms all baselines on all datasets, e.g., CLIMATES-I achieves a CV of 0.0989 on the RET dataset, whereas the next best-performing baseline achieved a CV of 0.1002. This establishes the superior performance of CLIMATES in providing accurate forecasts for AET, RET, and NPP. Additionally, we observe that the mentioned heuristic strategies for assigning data points to the clusters (i.e., CLIMATES-II and CLIMATES-III) lead to similar results. Note that although the improvement of CLIMATES over baselines does not look significant from an ML perspective, we will show, in the next section, that this improvement over baselines could result in considerable cost savings in the real world.

Orthogonally, Table 2.5 shows that although neural network models outperform popular statistical models by a relatively large margin, their performance is comparable to some strong classical ML models. This finding is consistent with prior research, as there is a growing body of work that questions the superiority of some recent neural networks over classical models. For example, this finding is consistent with results reported in prior work in the area of Recommendation Systems [47], which found that some recent neural network models are not actually superior to well-tuned classical models. As an analogous result in the time-series forecasting domain, our findings suggest that despite the easy availability of large-scale datasets in time-series forecasting (due to easy access to remote sensing data), deep learning

Model	AET	RET	NPP
TBATS	0.2414	0.1206	0.2856
SARIMA	0.2130	0.1029	0.2503
LR	0.2114	0.1014	0.2492
RF	0.2080	0.1022	0.2427
XGBoost	0.2099	0.1039	0.2445
SVM	0.2110	0.1041	0.2505
SFM	0.2080	0.1002	0.2428
TCN	0.2138	0.1012	0.2432
CLIMATES-I	0.2075	**0.0989**	**0.2409**
CLIMATES-II	**0.2071**	0.0990	**0.2409**
CLIMATES-III	**0.2071**	0.0990	**0.2409**

Table 2.5: CV of CLIMATES and various baselines

does not always beat classical ML models significantly.

2.6 Real-World Use Case

This section explains three possible ways that CLIMATES could be employed to help smallholder farmers in the field.

Application 1: Forecasting the Level of Water Stress. CLIMATES can be used to assist farmers in getting to know the occurrence of water stress in their farms ahead of time. Past literature suggests that water stress in each farm can be estimated from RET and AET using the following equation: $K_s = AET/(K_c \times RET)$ [39]. In this equation, K_s denotes the water stress index (e.g., $K_s < 0.5$ indicates an alarming level of water stress) and K_c refers to the crop coefficient, for which the suggested values are available at [39]. Thus, CLIMATES can serve as the engine of a mobile app that can send early warnings to farmers based on the future value of K_s computed from the forecasted AET and RET. Further, the output of CLIMATES can be used to generate a heatmap, similar to Figure 2.4, to represent the water stress index across a large region. In this figure, the background color shows the forecasted level of water stress (assuming $K_c = 1.2$) across Kenya on the first dekad of May 2019 and black circles represent particular

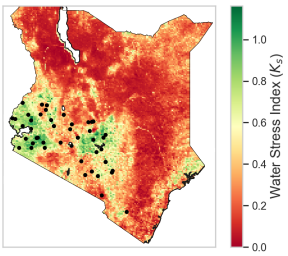


Figure 2.4: A heatmap of water stress index (K_s) in Kenya

farm locations. According to this heatmap, the farms in western Kenya are at low risk of water stress (as $K_s > 0.5$) during that particular dekad.

Application 2: Irrigation Scheduling. CLIMATES can be utilized as an AI assistant for irrigation scheduling as well. Irrigation scheduling methods aim to determine the timing of irrigation and the amount of water demand at different stages of the crop-growing life cycle. One common approach in this space is ET-based irrigation scheduling, which utilizes ET data to provide customized suggestions for each farm based on its irrigation system, crop type, etc. According to this approach, the amount of water demand can be estimated using Equation 2.2 [48]. In this equation, GI denotes the gross irrigation water requirement, ET_c denotes the crop evapotranspiration ($ET_c = K_c \times RET$), P_e refers to the effective precipitation that can be consumed by plants, and E denotes the efficiency of the irrigation system used in the target farm. Thus, providing farmers with information on the future value of GI (through forecasting RET) can help them estimate the amount of water needed for mitigating the water stress in their farms.

$$GI = \frac{ET_c - P_e}{E} = \frac{(K_c \times RET) - P_e}{E} \tag{2.2}$$

Real-World Impact of CLIMATES versus Baselines. We now compare the potential real-world impact of CLIMATES against the best-performing baseline model in the context of irrigation scheduling. To this end, we translate the amount of improvement in predictive accuracy (of CLIMATES over the best-performing baseline) to the corresponding difference between GI (i.e., required levels of irrigation) computed from the outputs of CLIMATES (i.e., $\text{GI}_{CLIMATES}$) and the best-performing baseline (i.e., $\text{GI}_{baseline}$). This difference in GI ($\text{GI}_{baseline}$-$\text{GI}_{CLIMATES}$) could be an indicator of the amount of water that could be saved as a

result of employing CLIMATES, rather than the best-performing baseline. However, translating the difference in predictive performance (in terms of CV) to the amount of water saving requires several assumptions as the value of CV does not distinguish under-estimation from over-estimation. In addition, the parameters of Equation 2.2 depend on various characteristics of the farm, e.g., K_c changes with the crop type and the stage of crop growth. For ease of exposition, we assume that $E = 0.60$ (which corresponds to the Surface irrigation system [49]) and $K_c = 1.2$ (which corresponds to mid-season maize cropping [39]) are used in the target region and that both CLIMATES and the best-performing baseline (i.e., SFM) overestimate RET on a given dekad in that region.

In this situation, according to Equation 2.3, the improvement of 0.0013 by CLIMATES over SFM in terms of CV in the RET prediction task (from Table 2.5) can be translated into saving about 92 liters of water per hectare each day for a maize-cropped farm at the mid-season stage. As a result, *although the improvement of CLIMATES against baselines looks small numerically, this improvement can result in considerable water saving when it comes to employing CLIMATES for scheduling irrigation within the crop growing season in the real world.*

$$GI_{baseline} - GI_{CLIMATES} = \frac{K_c}{E}(RET_{baseline} - RET_{CLIMATES}) \tag{2.3}$$

Application 3: Monitoring Crop Growth. CLIMATES can also be employed to quantitatively monitor plant growth. For example, NPP values forecasted by CLIMATES can be used to proactively identify some real-world stressors influencing plant growth such as nutrition shortage. In particular, CLIMATES can produce customized early warnings based on the amount of gap between the forecasted NPP and the NPP of the plant under non-stressed conditions.

2.7 Challenges in Implementation

There are several challenges that need to be taken into account when planning for deployment in this domain. First, many African smallholder farmers live in rural areas with limited access to the Internet, and CLIMATES is in need of frequent updates, and due to its computational needs, it needs to be updated on a server. Thus, access to the most recent information requires establishing a connection with

a server via the Internet, and consequently, cannot be done offline. To address this challenge, a feature could be added to the app for automatically sending frequent updates to the registered farmers via text messages (SMS) so that they can stay updated even in case of Internet connection issues. The second challenge is related to farmers' concerns about the privacy of their data. In fact, many farmers may not be willing to share some data such as farm size and crop type, as this information along with their estimated crop productivity could be used to derive their income, which is personal information to many people.

2.8 Summary

This chapter proposes CLIMATES, an ML-based meta algorithm for forecasting three important crop-productivity related variables (AET, RET, and NPP) in smallholder farms across Africa. Leveraging structural insights about these variables, it attempted to combine the power of several popular time-series forecasting techniques to produce more accurate forecasts in the face of significant variability, mainly stemming from the geographic and climatic diversity of different African countries. The experimental results show that CLIMATES outperforms several strong baselines, including VRNNs which introduce latent variables to model variability in time-series data.

Chapter 3
AI for Agriculture: Biotic Stress Prediction from Sparse Data

This chapter focuses on forecasting the presence/absence of a devastating biotic stressor, namely Desert Locust, from a diverse set of input data, in which some data is sparse, and not distributed uniformly across space/time [50]. In the following sections, we describe the problem domain, related work, our solution, experimental results, and the real-world use case that we envision for such a predictive model.

3.1 Introduction

In 2020, several parts of the world (especially East Africa and the Middle East) struggled with the worst Desert Locust (*Schistocerca gregaria*) swarm infestation in over 25 years [51]. The Desert Locust is a highly destructive pest during its swarming phase. Each 2g adult locust can move as much as 100 kilometers/day, consume its own weight in vegetation each day, and each swarm can contain billions of locusts [52, 53]. Additionally, the Desert Locust outbreak could have significant economic, human, and environmental impacts. For example, the 2020 Desert Locust crisis resulted in the forcible displacement of numerous people and the decimated crops left by these locust swarms jeopardized the food security of millions of people, particularly smallholder farmers [52]. Thus, it is critically important to accurately forecast their occurrence, so that appropriate mitigation measures can be taken.

To tackle this crisis, the Desert Locust Information Service (DLIS) at UN-FAO [53] has historically relied on highly trained staff conducting field surveys

in at-risk geographical areas, followed by governments allocating and spraying pesticides in affected regions. However, due to limited numbers of trained staff conducting field surveys, especially in countries where desert locusts are not normally present, the DLIS aims to augment its data collection and decision-making through crowdsourced data. As a result, in 2020, PlantVillage, at the request of UN-FAO, developed eLocust3m[1], a smartphone application that was designed for non-experts to use to crowdsource records of locust observations. The introduction of this eLocust3m application into a well-established system of surveillance by the DLIS offers opportunities to enhance current locust mitigation operations, particularly through ML-based approaches. Accordingly, this chapter builds a spatio-temporal ML model to forecast the Locust presence/absence with a high accuracy.

In fact, this chapter proposes PLAN (**P**redictor of **L**ocust **A**ctivity and moveme**N**t), a spatio-temporal deep neural network model that leverages real-world insights as well as the crowdsourced data of locust observations to accurately forecast locust presence/absence at high spatial and temporal resolutions across Kenya, Ethiopia, and Somalia (three countries in East Africa which have suffered great losses due to the Desert Locust crisis). In particular, in this chapter, we make the following main contributions: (1) through PlantVillage, a partner of the UN-FAO, we utilize data from eLocust3m (a first-of-its-kind tool that has been deployed in the field) and create an image representation of locust survey data to explicitly capture Locust movement patterns across space and time. (2) Leveraging subject matter expertise and findings of prior studies in the agriculture domain, we identify ten environmental factors that contribute to locust breeding, migration, and survival, and fetch remote-sensed data for each of these ten factors. (3) We propose PLAN, that takes as input a single geographical location (in terms of latitude and longitude) and outputs accurate n-day forecasts of locust presence/absence at that location through learning relevant features from different inputs. PLAN explicitly models the spatio-temporal relationships in locust movement, and the impact of environmental factors on locust movement using a combination of Convolutional Neural Network (CNN) [54] and Long Short-Term Memory (LSTM) [34], and Feed-Forward Neural Network (FNN) models. (4) Finally, we comprehensively evaluate the effectiveness of PLAN for this problem domain. The experimental results show that PLAN outperforms several classical ML baseline models (in terms

[1] `https://play.google.com/store/apps/details?id=plantvillage.locustsurvey`

of the predictive performance) on the nth-step ($n \in \{1, 2, 3, 4\}$) forecasting tasks. For example, PLAN is the only model which achieves an AUC score of $\sim$0.9 for next-day forecasts. More importantly, PLAN shows a significant improvement (23% higher F1 score) over the best-performing baseline model in a cross-region test (i.e., when we test the performance of ML models on the data of geographical regions which are far away from the regions where training data was collected), which illustrates PLAN's capability of learning useful locust migration patterns.

PLAN is meant to be an assistive tool, which can augment the human expertise of the highly trained staff at DLIS and PlantVillage in their locust prediction and mitigation efforts.

3.2 Related Work

Historically, locust swarm migration has been studied from multiple perspectives in prior work. One line of prior research focuses on exploring the role of climatic factors in the outbreak of migratory locust swarms [55, 56]. For example, various studies have reported that different meteorological factors can have different levels of impact on locust breeding, maturation, migration, and survival; e.g., (1) high precipitation can make a region suitable for locust breeding [56], similarly, (2) soil moisture was also found to be a strong indicator of locust breeding areas [57, 58], (3) wind can facilitate locust migration [56], (4) green vegetation plays a key role in locust survival [56], and finally, (5) increased temperature resulting from climate change tends to exacerbate the problem of locust swarm infestation [56].

In addition, few data-driven studies at the intersection of agriculture and AI have addressed the locust crisis. Ye et al. [59] employed CNN-based models to detect locust species from imagery data. Kimathi et al. [60] used the Maximum Entropy (MaxEnt) model to identify potential locust breeding spots from several environmental factors. Moreover, in January 2021, the Selina Wamucii company[2] announced the development of an AI tool (called *Kuzi*) for predicting locust occurrence and breeding, however, the details of their underlying model and its predictive performance were not released [61]. Therefore, to the best of our knowledge, there had been no prior publicly available research on forecasting locust presence/absence at high spatial and temporal resolutions. To fill this gap, this

[2] `https://www.selinawamucii.com`

chapter proposes PLAN, an ML algorithm that leverages recent advances in the field of spatio-temporal forecasting [62, 63], as well as findings of prior studies (in the agriculture domain) on locust outbreaks to generate accurate predictions.

3.3 Datasets

3.3.1 Raw Data Sources

PLAN utilizes two sources of raw data: (1) crowdsourced locust survey data; and (2) remote-sensed environmental data.

1. **Crowdsourced locust survey data.** This data is collected through the "eLocust3m" (or eL3m) Android application, which has been developed by PlantVillage for the UN-FAO. This smartphone application enables users to record observations of locust presence/absence at a particular geographical location (given by latitude and longitude) on a given date. Since 2020, eL3m has been deployed in many countries around the world, and various groups (such as PlantVillage, county governments, charities) have employed local community members to scout the areas and provide geocoded observations of locust presence/absence via eL3m. In this work, we mostly focus on eL3m locust presence/absence data collected from Kenya, Ethiopia, and Somalia (three countries which have been badly hit by the locust crisis) between March 1^{st}, 2020 to September 30^{th}, 2020. In total, during this time period, $\sim$21,000 locust presence/absence reports were recorded in these three countries via eL3m. We use all these reports as our first raw data source.

2. **Remote-sensed environmental data.** This source of data consists of some environmental factors that could affect locust breeding, migration, and survival. In fact, we take advantage of subject matter expertise and prior work in the agriculture discipline [56, 58], and fetch raw remote-sensed data for the following ten environmental factors from publicly available data sources cited below: (1) soil moisture [64–66], (2) sand content of soil [67], (3) precipitation [68], (4) land elevation [69], (5) wind speed at 10 meters [70], (6) wind speed at 50 meters [70], (7) U wind speed at 10 meters [71], (8) V wind

speed at 10 meters [71], (9) total biomass productivity in 2019 (TBP_19) [72], and (10) actual evapotranspiration (AET) [72].

Rationale for the Choice of Environmental Factors. Each environmental factor chosen by us has been reported in prior work as potentially having an impact on locust breeding, migration, or survival. For example, high sand content in soil, and soil moisture is conducive for locust egg-laying [56]; as a result, precipitation, soil moisture, sand content of soil, and AET can serve as strong indicators of potential locust breeding spots, which can assist in forecasting their presence/absence. Further, wind is regarded as the main means of locust migration [56]. The wind heights most important to locust movement are 1,000 and 1,500m above sea level. Here, we use wind speed at 10/50 meters and directions (i.e., U/V wind) as they were readily available. Finally, certain land characteristics are conducive to locust presence; e.g., high locust activity is seen at lower elevations [73], and green vegetation is needed for locust survival [56]. As a result, land elevation, and TBP_19 could play important roles in forecasting locust presence/absence in different regions.

3.3.2 Data Characteristics

Our eL3m data has certain characteristics, which mainly stem from the nature of crowdsourced data collection. First, as locust presence/absence is voluntarily reported by human eL3m users, the total number of reports received each day varies across time. Users often submit multiple records in close succession resulting in temporal and spatial aggregation. Figure 3.1 represents the total number of locust presence/absence reports received across Kenya, Ethiopia, and Somalia over time. As illustrated in this figure, a large number of locust presence/absence reports were received each day from the beginning of June until mid-July. In particular, most of the reports received in June are locust presence (or, positive) reports, whereas the majority of the reports received in July are locust absence (or, negative) reports. Second, the spatial distribution of the data is not uniform over time; e.g., on several days in June, there are many regions in Kenya from which no locust reports were received. Third, we acknowledge the presence of some noises in the data, because people voluntarily report the locust presence/absence, and they might not report

the ground truth intentionally/unintentionally, e.g., there are false positive reports where users have considered it important to submit positive records even if locusts are not present (we elaborate on this in Section 3.7).

In addition, remote-sensed environmental factors are available at different temporal resolutions. For example, wind speed, soil moisture, and precipitation are available at a daily resolution, whereas AET is only available at a dekadal resolution. On the other hand, sand content, TBP_19, and land elevation are static features that do not vary with time. In section 3.3.3, we describe our data preparation steps.

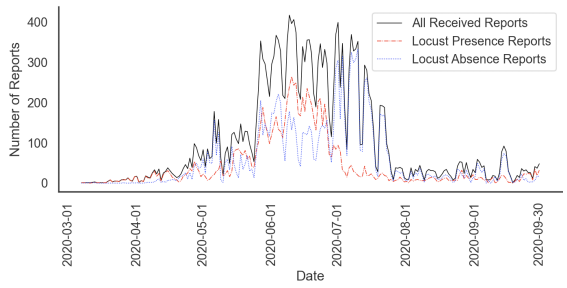


Figure 3.1: Distribution of eL3m locust presence/absence reports received from Ethiopia, Kenya, and Somalia over time

3.3.3 Data Preparation

In our dataset, each data point corresponds to a single eL3m locust report with a binary (present/absent) label. Each of these data points is recorded by an eL3m volunteer at a particular geographical location $(\text{lat}, \text{long})$ and date/time t. For example, Figure 3.2 illustrates all such data points recorded in Kenya on date t (similar maps can be drawn for different dates and countries).

In order to represent each individual data point in Figure 3.2 (without loss of generality, we denote an arbitrary point by the blue GPS pin), we create an image-based feature representation that can help summarize the non-uniformly distributed data and capture spatio-temporal relationships in the movement of locusts in nearby regions (surrounding the blue pin location) over the previous k days. In particular, for each of the previous k days, we create a separate $7 \times 7 \times 2$ image representation which summarizes all eL3m locust reports (both presence and absence) received from surrounding areas which lie in a 7×7 grid centered on the

blue pin location.

More formally, the feature representations for a data point corresponding to location l = (latitude, longitude) and date $(t+n)$ are created by generating k images of size $7 \times 7 \times 2$, one for each date $t' \in \{t, t-1, t-2, \ldots, t-(k-1)\}$. In order to build the image for date t', we grid the geographical area surrounding location l and create a 7×7 image, in which each pixel corresponds to a square geographical area of size $d^\circ \times d^\circ$ (in spatial resolution degrees). This image is centered on location l, hence the central pixel corresponds to a region of size $d^\circ \times d^\circ$ centered on location l, and other pixels correspond to nearby $d^\circ \times d^\circ$ regions. Finally, each pixel contains two pieces of information: (1) the total number of locust presence reports from that $d^\circ \times d^\circ$ region on date t', and (2) the total number of locust absence reports from that $d^\circ \times d^\circ$ region on date t'. Using this procedure, we create k images of size $(7 \times 7 \times 2)$.

Intuitively, this time-varying image representation of data points enables us to explicitly capture the movement of locusts across space and time which can serve as important predictors for future locust movement, e.g., locust presence in a region increases the likelihood of locust presence in nearby regions in the near future and vice versa. Thus, this set of k images (one for each of the previous k days) forms the first part of the feature representation for each data point in our dataset.

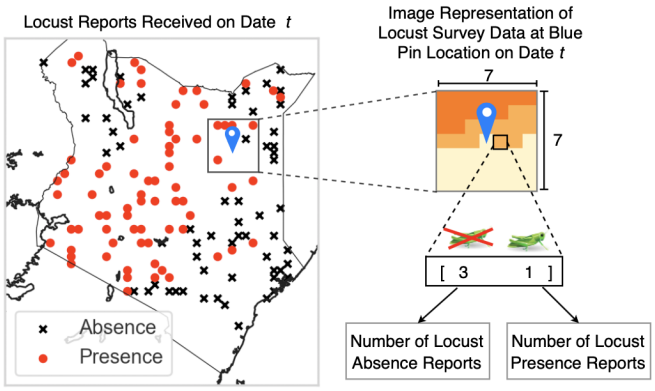


Figure 3.2: Schema for image representation of a single locust presence/absence report received on date t

The second part of feature representation for each data point comprises time-series values for six remote-sensed environmental variables (i.e., precipitation, soil moisture, U wind at 10 meters, V wind at 10 meters, wind speed at 10 meters, and

wind speed at 50 meters) over the previous k days. Finally, the third part of feature representation comprises single values for our static environmental variables (i.e., sand content, TBP_19, land elevation, and AET of the last dekad). We normalize each of these features independently via Min-Max normalization.

Our final dataset consists of 21,012 data points, out of which 42.35% correspond to locust presence reports (i.e., positive class). Each data point consists of the following input features: (1) k matrices of size $(7 \times 7 \times 2)$, which correspond to the image representation of locust survey data on each day of the past k days, (2) six time-series data of length k, each of which corresponds to the historical pattern of an environmental factor, and (3) a vector of four elements which corresponds to values of our four static variables.

3.4 The Proposed Framework: PLAN

We now describe PLAN, a deep learning framework for generating accurate forecasts of locust presence/absence at different geographical locations. PLAN takes the (latitude, longitude) of the target location and the current date t as input, and generates as output a binary forecast about whether locusts will be present (or not) at that (latitude, longitude) n days into the future (i.e., on the day $t + n$).

Figure 3.3 illustrates the network architecture of PLAN. At a high level, it consists of three components: (1) a CNN+LSTM network for capturing spatio-temporal relationships from our image-based feature representations; (2) an LSTM for capturing temporal relationships in time-series environmental variables; and (3) a Feed-Forward neural network (FNN) for extracting relevant features from the static environmental factors. In the following paragraphs, each component is explained in detail.

Module A: CNN + LSTM Model. We model spatio-temporal relationships in eL3m locust reports as follows. (1) For each data point, we build k image representations of locust report data (as described in Section 3.3.3) to summarize the locust reports received from surrounding regions over the last k days. (2) Each image is passed through a separate CNN network followed by a fully-connected (FC) layer, which outputs dense latent representations of the spatial relationships that exist in that image. (3) The output from each FC layer is then fed as input to the hidden state of an LSTM network which captures locust migration patterns

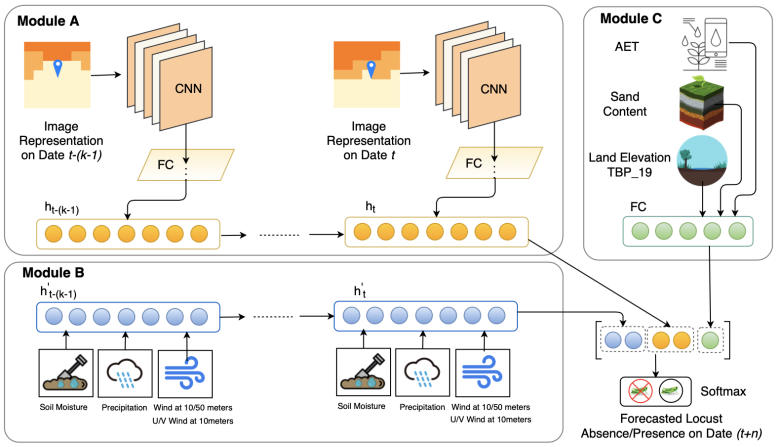


Figure 3.3: The architecture of PLAN

over space and time. Each of our CNN networks (in Figure 3.3) consists of three convolutional layers with 16 filters of size 3×3 and the FC layer has 64 neurons. Similarly, the hidden state size of the LSTM network is 256.

Module B: LSTM Model. We model the impact of environmental factors on locust movement as follows. (1) For each data point, we concatenate the time-series values of six environmental factors (i.e., soil moisture, precipitation, wind speed at 10/50 meters, and U/V wind at 10 meters) at that data point's geographical location over the previous k days. (2) This $6 \times k$ time-series data is passed through an LSTM network with h hidden states (with hidden state size = 64), which enables capturing time-varying patterns of environmental factors at a specific geographical location.

Module C: FNN Model. Prior studies in the agriculture domain show that locust presence could be associated with land characteristics. For example, sandy soil is favorable for locust breeding, and high locust activity is seen at lower elevations, etc. [56, 73]. As a result, PLAN takes four of such factors (i.e., land elevation, sand content of soil, TBP_19, and AET in the last dekad) as input and uses a FNN with a FC layer to extract relevant features of the target region from these factors.

Finally, the output representations discovered by the last hidden layers of LSTMs in Modules A and B as well as the output of Module C are fed into a softmax layer to generate a predicted forecast of locust presence/absence n days into the future.

3.5 Experimental Evaluation

In this section, first, we discuss our evaluation approach and experimental settings. Then, we evaluate PLAN's performance as follows: (1) We compare PLAN with several baseline models to show its superior predictive performance on various forecasting tasks. (2) We conduct an ablation analysis to show the impact of different parts of PLAN's architecture on its predictive performance. (3) We conduct a cross-region test to evaluate PLAN's performance when being tested on the data of a distant geographical region that is far away from the training region. (4) To tackle data sparsity, which stems from the unavailability of reports from many geographic regions, we propose a model-agnostic data augmentation algorithm, and then, assess its effectiveness in this problem domain.

3.5.1 Evaluation Approach

To evaluate the performance of various models, we take advantage of the walk-forward testing approach [74–76] and adapt it to our problem domain. At a high level, walk-forward testing extends the idea behind K-fold cross-validation to sequential time-series data. This method enables a more robust and trustworthy assessment of the performance of various forecasting models because each model is evaluated under a series of time-varying conditions. In our domain, given the heterogeneous distribution of eL3m locust reports over time (see Figure 3.1), walk-forward testing enables us to evaluate our models' performance across a number of sequenced and time-shifted train/test splits.

However, in walk-forward testing, the overall predictive performance of a forecasting model (in terms of F1) is computed by averaging the F1 score achieved by the model across different time-shifted test sets (similar to macro-averaging in multi-class classification). Unfortunately, in our problem domain, the total number of eL3m locust reports per day changes over time (see Figure 3.1). Consequently, the different time-shifted test sets created during walk-forward testing have different numbers of data points. Therefore, it is not fair to report the average F1 score (or other evaluation metrics) computed on each time-shifted test set as the overall predictive performance of the model. To address this challenge, we combine all time-shifted test sets (and the predictions on those test sets) into a single larger

test set. We compute all evaluation metrics on this single test set, and use these metrics to evaluate and compare different forecasting models. Prior literature has shown that this approach to computing the overall performance produces unbiased estimates of predictive performance in several situations, e.g., this approach is commonly used with k-fold cross-validation, etc. [77].

3.5.2 Set-Up

All experiments are run on a machine with one NVIDIA Tesla T4 GPU, 4 vCPUs, and 15 GB RAM. Except for Table 3.1, the window length w of walk-forward testing is set to 7 days in all experiments. Finally, all experiments are run five times, and the average performance over all runs is reported.

The hyper-parameters are set as follows. All fully connected layers in PLAN's architecture use Sigmoid activation function. The batch size is set to 64, and the Adam optimizer [78] with a learning rate of 0.001, β_1 of 0.9, and β_2 of 0.999 is used. For all our experiments, we set the value of $k = 7$, i.e., both LSTM networks in Modules A and B of PLAN's architecture (Figure 3.3) take the data of the past 7 days as input. Further, we set the value of $d = 0.2°$, i.e., the spatial resolution of each pixel in our image representations (Figure 3.2) is set to 0.2°, which makes each pixel correspond to a 22.2 km $\times$ 22.2 km geographical region. The value of d was set via hyperparameter tuning.

3.5.3 Comparison with Baseline Models

To evaluate the effectiveness of PLAN, we compare its performance with the following baselines: (1) Logistic Regression (Logit), (2) Support-Vector Machine (SVM) with RBF kernel [43], (3) AdaBoost [79], and (4) XGBoost [42]. Building these baseline models requires one further pre-processing step as they cannot handle imagery data; i.e., we flatten the image representations of the eL3m locust report data, and concatenate them with all environmental factors to build the input feature representations for these baseline models. *Note that we chose these classical ML models as baselines, as there was no comparable prior work on sophisticated deep learning models to predict locust movement. Thus, any deep learning model that we compare PLAN against would have to be developed from scratch. Further, we note that during our ablation analysis, we compared PLAN against several*

	$w = 7$ days			$w = 14$ days			$w = 21$ days		
Model	Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC
Logit	0.7417	0.7026	0.7810	0.7366	0.6958	0.7652	0.7233	0.6823	0.7398
SVM	0.7772	0.7303	0.8433	0.7678	0.7104	0.8076	0.6853	0.5545	0.7870
AdaBoost	0.7585	0.7317	0.8282	0.7492	0.7247	0.8137	0.7408	0.7049	0.8002
XGBoost	0.7848*	0.7612*	0.8650*	0.7730*	0.7436*	0.8493*	0.7516*	0.7206*	0.8338*
PLAN	**0.8174**	**0.7918**	**0.8904**	**0.8060**	**0.7750**	**0.8781**	**0.8052**	**0.7814**	**0.8798**
Improv.	+4.15%	+4.01%	+2.93%	+4.26%	+4.22%	+3.39%	+7.13%	+8.43%	+5.51%

Table 3.1: The predictive performance of different ML models on the 1^{st}-step prediction task with various window lengths (w)

neural network architectures (i.e., variants of PLAN) to evaluate the contribution of different modules to PLAN's performance.

Table 3.1 compares the predictive performance of PLAN against baseline models on 1^{st}-step prediction tasks (i.e., next day forecasts) with different choices of window lengths $w \in \{7, 14, 21\}$ (for walk-forward testing). The best model's performance is shown in bold, whereas the second-best model's performance is shown with an asterisk. According to the results, PLAN consistently outperforms all baseline models; in particular, PLAN achieves an F1 score of ~ 0.792 with a window length $w = 7$, which improves upon XGBoost's (the best-performing baseline) performance by $\sim 4\%$. *This is a significant finding, as the high-stakes nature of decision-making in this domain means that any increases in predictive accuracy over baseline models could potentially lead to widespread impact (in terms of increased food security, better management of the locust crisis, etc.) at the scale of nations.*

In addition, PLAN tends to be more robust to increasing window lengths w as compared to baseline models. In particular, the percentage improvement achieved by PLAN over XGBoost (in terms of F1) significantly increases with increasing window length sizes. For example, PLAN improves upon XGBoost's F1 score by 4.01%, 4.22%, and 8.43% with window length sizes of w =7, 14 and 21 days, respectively. This finding illustrates that with increases in the window length size w, the distribution of training and test sets are more likely to differ from each other; as a result, the performance of all models is likely to degrade. However, Table 3.1 shows that PLAN is significantly less sensitive to potential covariate shift problems as compared to baseline models, e.g., when the window length is increased from $w = 7$ to $w = 21$, PLAN's F1 score minimally degrades by $\sim 1\%$, whereas XGBoost's F1 score degrades by $\sim$5%.

	2^{nd}-step prediction			3^{rd}-step prediction			4^{th}-step prediction		
Model	Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC
Logit	0.7297	0.6933	0.7691	0.7210	0.6907	0.7692	0.7166	0.6855	0.7634
SVM	0.7569	0.7050	0.8244	0.7424	0.6887	0.8114	0.7407	0.6956	0.8017
AdaBoost	0.7438	0.7158	0.8028	0.7303	0.7014	0.7834	0.7302	0.7049	0.7992
XGBoost	0.7717	0.7522	0.8507	0.7578	0.7374	0.8346	0.7507	0.7329	0.8380
PLAN	**0.7908**	**0.7588**	**0.8637**	**0.7726**	**0.7429**	**0.8497**	**0.7692**	**0.7340**	**0.8427**

Table 3.2: The predictive performance of different ML models for 2^{nd}-step, 3^{rd}-step, and 4^{th}-step prediction tasks

Next, we evaluate the predictive performance of different models on the nth-step prediction task, as forecasting farther ahead into the future tends to be a more difficult task. Table 3.2 compares the predictive performance of different models on 2^{nd}-step, 3^{rd}-step, and 4^{th}-step forecasting tasks. As expected, the performance of all ML models degrades with increasing forecast horizons. However, PLAN consistently outperforms baseline models at all forecast horizons. In particular, PLAN achieves an average AUC of 0.85 (across all horizon values) which shows its high capability of distinguishing positive/negative samples even when forecasting farther ahead into the future. *In summary, Table 3.1 establishes PLAN's superior performance against strong classical ML baseline models on a real-world task for which no comparable prior deep learning models existed.*

3.5.4 Ablation Study

Having established PLAN's superior performance, we now conduct two sets of ablation studies to investigate the impact of different parts of PLAN's architecture on its overall performance. Our first ablation study evaluates the impact of different input features on PLAN's performance. We build the following variants of PLAN: (1) PLAN\Env: All ten input environmental variables (both time-series and static ones) along with Modules B and C are removed from PLAN's architecture. (2) PLAN\eL3m: All eL3m locust report data along with Module A is removed from PLAN's architecture. (3) PLAN\LAbs: Instead of using dual-channel image representations of eL3m locust reports (where we store both the numbers of locust presence and absences reported at each pixel), we experiment with single-channel image representations by only storing locust presence numbers at each pixel in the image; as a result, the size of our input images becomes $7 \times 7 \times 1$.

Our second ablation study investigates the impact of different components of PLAN's architecture on its predictive performance (i.e., all input features are used for the prediction task, but the architecture is changed). We build the following variants of PLAN: (1) PLAN\CNN: CNNs are removed from the architecture of PLAN; instead, all image data is flattened and is passed through the FC layers in Module A, the output of these FC layers is passed into the LSTM network in Module A. Modules B and C are unchanged in PLAN\CNN. (2) PLAN\LSTM: Both LSTMs are removed from the architecture of PLAN; instead, the outputs of FC layers in Module A, the inputs of Module B, and the output of Module C are concatenated and fed into the output layer. (3) PLAN\CNLS: All LSTMs and CNNs are removed from the architecture, and instead, a FC layer with the same number of neurons as the size of the LSTM hidden state is used to replace those networks. Therefore, PLAN\CNLS is similar to a Multi-Layer Perceptron model.

Table 3.3 compares the predictive performance of our different ablations on the 1^{st}-step forecasting task. The results show that PLAN\eL3m (which ignores eL3m data along with Module A) leads to the greatest decrease in PLAN's predictive performance by reducing F1 scores by $\sim$22%. This illustrates the importance of the crowdsourced eL3m data in the predictive performance of PLAN. Further, PLAN\LAbs, which removes locust absence information from the input images (of Figure 3.2) results in a 6.97% decrease in F1 score, which shows that locust presence reports (by themselves) are not enough to generate accurate forecasts, and incorporating locust absence reports in image-based feature representations has a significant impact on the performance of PLAN. Additionally, PLAN\Env, which removes environmental factors, results in a 1.47% decrease in F1 score, which is consistent with domain insights on the role of environmental factors in locust activity and movement. Results from our second ablation study show that removing CNNs (i.e., PLAN\CNN) or LSTMs (i.e., PLAN\LSTM) from the architecture leads to $\sim$1.2% reduction in F1 score. Additionally, removing both CNNs and LSTMs results in further decrease ($\sim$2.6%), in F1 score. This shows that different components of PLAN play roles of differing importance in its overall predictive performance.

Model	Accuracy	F1	AUC
PLAN\Env	0.8057	0.7801	0.8822
PLAN\eL3m	0.6381	0.6162	0.7039
PLAN\LAbs	0.7398	0.7366	0.8512
PLAN\CNN	0.8050	0.7822	0.8816
PLAN\LSTM	0.8109	0.7820	0.8835
PLAN\CNLS	0.7960	0.7708	0.8686
PLAN	**0.8174**	**0.7918**	**0.8904**

Table 3.3: The results of ablation study

3.5.5 Cross-Region Test

Until now, we trained PLAN on a portion of the data collected from Kenya, Ethiopia, and Somalia, and tested it on another portion of the same data. Now, we evaluate the performance of PLAN when trained and tested on datasets from two geographically distant regions. We hypothesize that in this cross-region test, Module A, (i.e., the component designed for capturing spatio-temporal patterns of locust movement) should still be able to learn useful location-agnostic patterns of locust migration.

For this purpose, in addition to the data from Kenya, Ethiopia, and Somalia, we use eL3m data collected from Iran during the same time-period (i.e., March 1, 2020 to September 30, 2020) which consists of 5,117 locust reports. To check the aforementioned hypothesis, in each iteration of walk-forward validation, we use the same training portion of the data from Kenya, Ethiopia, and Somalia to train the PLAN model. Then, we replace the test data with the locust reports received from Iran in that particular test period and evaluate the performance of the trained model on this new test set.

Table 3.4 shows the predictive performance achieved by PLAN and XGBoost in our cross-region test on the 1^{st}-step prediction task. As expected, the predictive performance of both ML models degrades in this cross-region test. However, PLAN consistently outperforms XGBoost on each evaluation metric, e.g., PLAN achieves $\sim$23% higher F1 score than XGBoost in this cross-region test. More importantly, comparing PLAN with PLAN\Env shows that removing environmental factors from PLAN results in a significant improvement in its predictive performance when being tested on the data of Iran. This improvement (that results from removing

Model	Accuracy	F1	AUC
XGBoost	0.5276	0.3322	0.6464
PLAN	0.6576	0.4115	0.7480
PLAN\Env	**0.7363**	**0.4819**	**0.8062**

Table 3.4: The results of cross-region test (i.e., the models are trained on the data of three East African countries and tested on the data of Iran)

environmental factors) makes sense because the climatic conditions in Iran differ completely from conditions in Kenya, Ethiopia, and Somalia. Consequently, training our models on environmental variables from East Africa could add noise to the model's forecasts when tested on Iran. Additionally, PLAN\Env achieves an AUC of $\sim$0.8, which indicates its capability in learning useful locust movement patterns that can help it generate relatively accurate forecasts about locust presence in regions located far away from the training region.

3.5.6 Model-Agnostic Data Augmentation

As locust observations were voluntarily reported by human eL3m users, locust reports are not available for many geographical regions on any given day. Thus, there are many 0's in the image representations of eL3m locust report data, as each image summarizes the total number of locust (presence/absence) reports received from a specific region on a particular day. To account for this data sparsity, we implement a model-agnostic linear interpolation approach for data augmentation and evaluate its impact on model predictive performance.

Our linear interpolation-based data augmentation approach relies on the following intuition about locust movement: if locusts are reported to be present (absent) in location l on two separate days (t_1 and t_2) that are close in time, it is highly likely that locusts are present (absent) at location l on all the days between t_1 and t_2.

More formally, in our data augmentation procedure, to forecast locust presence/absence in location l on date t, we take the following steps after creating image representations of eL3m reports received by date $(t-1)$: (1) if no locust report (neither locust presence nor locust absence) is available for a specific region, we set the value of the corresponding pixel to NULL in both image channels (locust

	Accuracy			F1			AUC		
Model	Before	After	Gain (%)	Before	After	Gain (%)	Before	After	Gain (%)
Logit	0.7417	0.7709	+3.93%	0.7026	0.7303	+3.94%	0.7810	0.8130	+4.09%
SVM	0.7772	0.7859	+1.11%	0.7303	0.7259	-0.60%	0.8433	0.8544	+1.31%
AdaBoost	0.7585	0.7978	+5.18%	0.7317	0.7608	+3.97%	0.8282	0.8651	+4.45%
XGBoost	0.7848	0.8099	+3.19%	0.7612	0.7749	+1.79%	0.8650	0.8819	+1.95%
PLAN	0.8174	0.8306	+1.61%	0.7918	0.8036	+1.49%	0.8904	0.9021	+1.31%
Avg			+3.00%			+2.11%			+2.62%

Table 3.5: Impact of data augmentation on the predictive performance of different ML models

presence and absence channels of the image). (2) For each region (i.e., pixel), we impute the time-series data of locust presence (absence) at each pixel separately using linear interpolation. (3) If no reports are available from specific regions, all elements of the time-series data could be NULL. Therefore, the remaining NULL values are replaced with 0 again. This procedure enables us to impute values for pixels that contain $[0, 0]$ (i.e., pixels that have no locust presence and absence reports at all). Further, to forecast locust presence/absence in location l on date t, we do not rely on the reports received after date $(t-1)$, and therefore, this data augmentation approach is consistent with the time-series nature of the problem.

Table 3.5 shows the impact of data augmentation on the predictive performance of ML models on the 1^{st}-step forecasting task. Each evaluation metric's value before/after data augmentation is reported in the Before/After columns, respectively. The percentage of improvement achieved by applying data augmentation is reported in the Gain column. Table 3.5 shows that incorporating data augmentation improves the predictive performance of all ML models; in particular, it improves the accuracy and F1 score by about 3.0% and 2.1%, respectively (on average), which shows this data augmentation technique's effectiveness in this domain. Importantly, PLAN achieves an AUC of $\sim$0.9 with this data augmentation technique. Thus, we propose to use PLAN with this data augmentation technique in future operational systems.

3.6 Real-world Use Case

One possible way in which PLAN can be used to assist farmers, policymakers, and human experts at UN-FAO is through the generation of high-resolution heatmaps (containing accurate forecasts of locust presence/absence). These heatmaps can

give all three stakeholders an improved understanding of the future susceptibility of locust swarm infestation for different geographical regions, which in turn, can hopefully help them make a more well-informed locust mitigation plan. For example, these heatmaps can assist decision-makers in strategically allocating scarce resources (e.g., helicopters, pesticides, etc.) among high-risk geographical areas in order to ensure efficient resource usage and a corresponding reduction in locust populations.

Figure 3.4 illustrates a heatmap of 1^{st} step forecasts (for June 10^{th}, 2020) generated by PLAN. This heatmap is generated by running PLAN's prediction model for each geographical location in Kenya on June 10^{th}, 2020. This heatmap shows North West Kenya and East Kenya as two potential hotspots of locust presence on June 10^{th} (characterized by a high predicted likelihood of locust presence), whereas it shows Central Kenya as a potential source of locust absence reports (characterized by low predicted likelihood of locust presence). In this figure, the light blue circles and light green crosses show the eL3m locust presence and absence reports (respectively) received on June 10^{th}, 2020 across Kenya. These circles align well with our forecasted hotspot in North West Kenya, whereas the crosses align well with Central Kenya. Thus, this indicates that PLAN's predictions have high recall in this example.

In order to understand why PLAN forecasted East Kenya as another hotspot, we plot dark blue circles and dark green crosses to represent eL3m locust presence and absence reports (respectively) received from June 11^{th} to 13^{th}, 2020 across Kenya. Interestingly, the dark blue circles align extremely well with the forecasted hotspot in East Kenya, whereas most of the dark green crosses align well with Central Kenya. We hypothesize that this is due to delays in data reporting by human volunteers, i.e., locusts arrived in East Kenya on June 10^{th}, but they were reported by eL3m users on June 11^{th} to 13^{th}. Since we don't have ground truth information, it is impossible to completely validate this hypothesis. However, we argue that the forecasted hotspot in East Kenya should not be viewed as false positives output by PLAN, as eL3m locust presence reports are recorded from the East Kenya hotspot within a period of 24 hours of our day of forecast. This illustrates that PLAN's predictions also possibly have high precision.

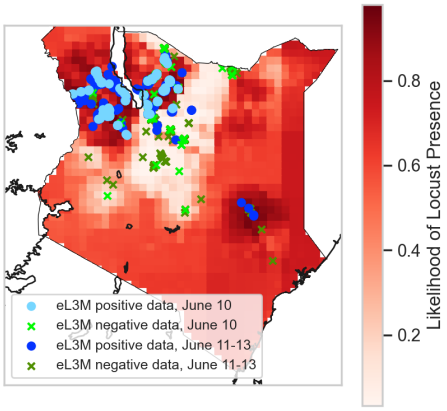


Figure 3.4: PLAN's forecasts about the likelihood of locust presence across Kenya on June 10^{th}, 2020 along with the ground truth reports received from Kenya on this particular date

3.7 Challenges in Implementation

The ubiquity of smartphones offers the possibility that well-designed mobile apps (such as eL3m) can enable the collection of large amounts of data in a short period of time. For humanitarian challenges like locusts (but also including floods, droughts, and other pests that damage crops), the potential benefits of such an application are very high. However, the major trade-off is data quality. Here we sought to use the data received from the crowd as-is in order to develop an ML model that could effectively use noisy data. We found that PLAN has led to an increased predictive performance over baseline models. While this is recognized, we understand that a major implementation challenge is the acceptance of such approaches by local actors such as governments in charge of the control operations. It would be expensive in both resources and time to deploy control operations to areas where locusts do not generally occur, but the model predicts their presence. As such, we think a major challenge will be to familiarise the decision-makers with the opportunities and pitfalls associated with ML-augmented desert locust predictions. We think one important role that PLAN could play is helping human experts more readily spot false records submitted by the crowd. This would reduce time spent in cleaning up databases which is currently a major task for staff at DLIS and PlantVillage. Thus, we hope that the use of PLAN would lead to a greater acceptance of ML to augment the human expertise at PlantVillage, UN-FAO, and other stakeholders.

3.8 Summary

This chapter proposes PLAN, which relies on a modular neural network architecture to forecast the locust presence/absence from crowdsourced data as well as remote-sensed environmental data. Experimental results show that PLAN achieves a superior predictive performance against several classical ML baseline models on a wide variety of forecasting tasks.

Chapter 4
AI for Social Welfare of Housing-Insecure Low-Income Americans: Eviction Filing Prediction with Fine-Grained Ground-Truth Labels

This chapter focuses on the eviction crisis faced by many low-income renters in the United States and develops predictive ML models to forecast the number of tenants at-risk of formal eviction when fine-grained ground-truth labels are available [80]. In the following sections, we describe the problem domain, related work, our solution, experimental results, and the real-world use case that we envision for such a predictive algorithm.

4.1 Introduction

Eviction is an urgent societal issue, which severely affects the lives of low-income individuals in the U.S. from multiple perspectives. In particular, it puts evicted families into material hardship [20] and could increase the risk of various health issues (such as depression and parental stress) and reduce their prospects of future decent housing [20, 22, 81, 82]. Furthermore, it could intensify various types of social problems such as poverty and housing inequality [82, 83]. Therefore, tackling

the eviction crisis plays a critical role in improving the lives of this vulnerable population, and helps make a progress on the SDG #1 "No Poverty" and SDG #11 "Sustainable Cities and Communities"[1].

To mitigate the eviction crisis, several eviction prevention/diversion programs (such as the Emergency Rental Assistance Program[2]) have been designed and implemented in the field. In particular, the federal government has allocated various financial resources (such as cash assistance, vouchers, etc.) to help households who have difficulty paying their rent. In spite of their availability nationwide, there is a large variability in the use of those resources; i.e., while the allocated resources have been used completely in some regions, parts of the allocated resources have been returned to the federal government from some other regions [84]. This observation suggests a need for a more efficient resource allocation strategy, which in turn, requires more accurate forecasts of the future number of tenants at-risk of eviction in target regions. Thus, any attempt to improve the accuracy of the forecasted number of tenants at-risk of eviction could have substantial impacts on the effectiveness of existing policies to disperse resources.

To this end, this chapter leverages recent advances in the ML domain [50, 62] to forecast the number of tenants at-risk of formal eviction in various census tracts[3] at a temporal resolution of one month. Our model, named as MARTIAN (Multi-view model forcAsting the numbeR of Tenants at-rIsk of formAl evictioN) leverages data sources of various spatial and temporal resolutions (namely, eviction filing records, American Community Survey (ACS) data[4], and labor statistics) to forecast the total number of tenants at-risk of eviction in each census tract n months into the future. Then, we evaluate the predictive performance of MARTIAN under various conditions using a real-world dataset consisting of information about eviction cases filed across Dallas County, TX since 2019. The results of our experiments show that MARTIAN outperforms a wide variety of baseline models in all considered situations; in particular, it achieves 5% lower Root Mean Square Error (RMSE) than the best-performing baseline model (on average). Further, it achieves a Spearman of 0.685, which shows that the ranking of census tracts is preserved to a

[1] `https://sdgs.un.org/goals`

[2] `http://tiny.cc/3vhouz`

[3] A census tract is a sub-region of a county and is defined by the U.S. Census Bureau for taking surveys and representing its results.

[4] `https://www.census.gov/programs-surveys/acs/data.html`

high extent in MARTIAN's forecasts. Additionally, the results of our cross-region test suggest that MARTIAN's superior predictive performance is generalizable to unseen census tracts. This research has been conducted in collaboration with Child Poverty Action Lab (CPAL)[5], which is an NGO aiming at tackling poverty-related issues across Dallas County, TX.

4.2 Related Work

As a pathway into various social problems (such as homelessness) [85,86], the eviction crisis has drawn the attention of scholars from several disciplines. In particular, there has been extensive research in social science literature on understanding the risk factors[6] of eviction and its consequences. As a result, past literature found three key categories of risk factors: (1) individual-level factors such as the number of children, job loss, and drug use disorder [87–90]. (2) neighborhood-level factors such as crime rate, and eviction rate in a neighborhood [88]. (3) network-level factors such as the number of disadvantaged people in a tenant's network [88].

Additionally, there has been a growing body of knowledge on the consequences of eviction and its impacts on individuals' lives. For example, prior work found that eviction could result in various health issues such as parental stress and depression [20, 91]. Furthermore, once getting evicted, tenants' credit rating gets debased, which in turn, puts more distance between them and the public housing program, and could exacerbate the housing inequality in society [92]. Although these empirical findings are informative and conducive to understanding the whole context of eviction, these studies do not focus on the problem of forecasting the number of tenants at risk of eviction. In contrast, this chapter leverages the findings of prior work in social sciences as well as ML techniques to forecast the number of tenants at risk of eviction, which could assist the government and NGOs in proactively tackling the eviction crisis in a more efficient and effective manner.

In addition to the social science studies, there has been some research from the AI community on mitigating the housing problems. For example, Ye et al. [93] and Tan [94] employed ML techniques to predict the risk of landlord harassment and the eviction rate, respectively. However, these studies have some limitations: (1)

[5] `https://childpovertyactionlab.org`

[6] Risk factors denote factors that are linked to the higher chance of a negative outcome.

the developed ML models forecast at the temporal resolution of one year, which limits their usability in our problem domain, where a forecasting tool with a higher temporal resolution (such as one month) is needed, or (2) they mainly relied on classical ML models and did not consider differences in the nature of various data sources in their design, e.g., time-series data and static data are treated the same. To address these limitations, we build a deep learning-based model that leverages various data sources with different spatial and temporal resolutions to forecast the number of tenants at-risk of getting formally evicted at the monthly resolution. Further, we conduct extensive experiments under various conditions to assess the superiority of MARTIAN to a wide variety of baseline models.

4.3 A Problem Statement

This chapter aims at building an ML model to precisely forecast the number of tenants at-risk of formal eviction (i.e., the number of eviction filings) at each census tract n months into the future.

Assume that E_t^c refers to the total eviction cases filed at census tract c in month t and L_t^c is a vector of length q representing the labor statistics at census tract c in month t (q refers to the total number of features in the labor statistics data). Also, suppose that ACS_t^c is a vector of length r representing the most recent values of ACS factors available at month t for census tract c (r refers to the total number of features selected from the ACS data). Note that as the U.S. Census Bureau releases the ACS data with a delay of about two years, at each point of time, ACS_t^c contains statistics of two years ago. Then, this chapter aims at building a forecasting model M such that:

$$M : E_{t+n}^c \leftarrow f(\{E_{t-k+1}^c, ..., E_{t-1}^c, E_t^c\}, \qquad \{L_{t-k+1}^c, ..., L_{t-1}^c, L_t^c\}, ACS_t^c)$$

Please note that the value of k is chosen through hyper-parameter tuning, and a separate experiment has been conducted for different values of n.

Data Source	An Explanation of Selected Input Feature(s)
Eviction Records	Historical data on the total number of eviction cases filed in each census tract

Labor Statistics	Unemployment rate
	Historical data on the number of employees in each of the following non-farm industries[7]:
	`Mining, Logging and Construction` – `Education&Health Services` – `Manufacturing`
	`Information` – `Leisure&Hospitality` – `Professional&Business Services` – `Government`
	`Trade, Transportation, and Utilities` – `Financial Activities` `Other Services`

ACS	# of renter-inhabited units # of renter-inhabited housing units, for which % of income contributing to housing expenses $\geq 30\%$ # of renter-inhabited housing units, for which the householder's income ≤ 0 in the last 12 months # of families receiving SSI and/or cash public assistance income who are below the poverty level # of renter-inhabited housing units, for which the householder's literacy level < high school # of renter-inhabited housing units, for which the householder's literacy level = high school graduate # of renter-inhabited housing units, for which the householder's literacy level = a college or associate's degree # of renter-inhabited housing units, for which the householder's literacy level = bachelor's degree or higher

Table 4.1: The definition of input features.

4.4 Datasets

This work uses three data sources: (1) Eviction filing records, (2) Labor statistics, and (3) American Community Survey (ACS). We extract our input features using these data sources, which are then used by an ML model to compute the value of the target variable. Table 4.1 provides detailed information on the input features extracted from each data source. In the following paragraphs, we introduce each data source and explain why we incorporate them into the model.

(1) Eviction Filing Records. This dataset consists of detailed information about eviction cases filed in judicial courts across Dallas County, TX. We get access to this dataset via CPAL, which receives daily updates (except for holidays) on new eviction cases filed in Dallas County. Each eviction record contains detailed information, e.g., the plaintiff's name, the defendant's name and address (geographical coordinates), the filing date, etc. However, the court's final decision regarding each case is not available in our dataset.

In this work, we use eviction filing data (since 2019) to compute the target variable and extract input features; in particular, we use the historical data on the number of eviction filings as input because overall eviction rate in a neighborhood is found to be associated with a greater likelihood of individuals' eviction [88].

(2) Labor Statistics. The U.S. Bureau of Labor Statistics releases monthly data on labor statistics[7], which contains various pieces of information related to the

[7] `https://www.bls.gov/eag/eag.tx.htm#eag_tx.f.2`

economy of a region, e.g., the unemployment rate and the number of employees in various non-farm industries (e.g., manufacturing and government). This data enables policymakers to monitor the economic/employment status over time and to make appropriate policies accordingly. Given a strong association between work status and the risk of eviction [88, 90], we believe that this data would provide useful signals to MARTIAN regarding monthly work status. This data is mainly released for each metropolitan area (rather than each census tract) and we use the data of the "Dallas-Fort Worth-Arlington" area.

(3) American Community Survey. The U.S. Census Bureau releases the ACS data, which is basically an annual report on various demographic/housing characteristics of different regions across the U.S. In particular, for renter-inhabited housing units (and their householders), it summarizes the value of the following metrics, which are found to have some associations with the risk (or number) of eviction and housing instability: work status [88, 90, 95], educational attainment [90, 96], income level, and monthly housing cost per income [21]. Accordingly, we utilize the 5-Year Experimental Estimates ACS data, which is available for each census tract in our study. Although the ACS data is not available at the monthly resolution, we think that it could still provide an insightful big picture of the situation in various census tracts.

Pre-processing. To pre-process our data, we take three main steps: (1) similar to prior work [97], we remove eviction filing records with commercial defendants and duplicate records from the dataset of eviction filing records, (2) we compute the total number of eviction filings in each census tract (out of 529 census tracts within Dallas County, TX) per month, (3) we scale the data of each input feature and the target variable into the range of [0, 1] using the Min-Max normalization. Please note that the predictive performance of all models is calculated after transforming the data to the original range (the parameters of min-max normalization are computed using the training data).

4.5 The Forecasting Model: MARTIAN

In this section, we explain our forecasting model. Leveraging recent advances in the ML domain [50, 62], we build a multi-view neural network to incorporate data

sources of different spatial/temporal resolutions into the prediction process. Figure 4.1 represents the architecture of MARTIAN. As illustrated, it has three views, each of which extracts features from one of the three aforementioned data sources: (1) The first view employs a Long Short-Term Memory (LSTM) network [34] followed by two fully-connected layers to learn patterns from the time-series data of eviction filings in the census tract of interest, (2) the second view extracts features from the time-series data of labor statistics using an LSTM network followed by two fully-connected layers, and (3) the third view employs a Multi-Layer Perceptron (MLP) to learn features from the factors selected from the ACS data. Then, the outputs of these three views are concatenated and given to the output layer to forecast the value of the target variable.

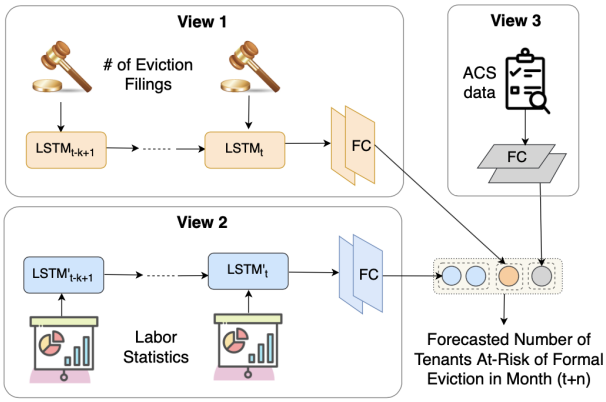


Figure 4.1: The architecture of MARTIAN.

4.6 Experimental Evaluation

In this section, first, we explain our experimental set-up and baseline models. Then, we compare the predictive performance of MARTIAN with that of baseline models and conduct an ablation study. Finally, we conduct a cross-region test to evaluate its generalizability.

4.6.1 Set-Up

To have a trustworthy and robust evaluation of the predictive accuracy of different models, we employ the walk-forward testing approach with a window length (w)

corresponding to 3 months [74]. Then, for each performance metric, we compute and report the average performance over all test sets. In addition, to train neural network models, the batch size, loss function, and maximum number of epochs are set to 32, MSE, and 200, respectively. We also utilize an Adam optimizer [98] with a learning rate of 2×10^{-4} and the early stopping approach [99] with a patience value of 10 epochs. Finally, as a result of hyper-parameter tuning, the value of k (i.e., the length of time-series inputs) is set to 6.

4.6.2 Comparison with Baseline Models

We compare the predictive performance of MARTIAN with that of an extensive set of baselines. The first set of baselines consists of the following classical ML models: (1) Ridge regression [100] (2) Support-Vector Machine (SVM) [43], (3) XGBoost [42], (4) Random Forest [41], and (5) LightGBM [101]. Additionally, we considered various deep learning-based models in our study as well. In particular, we conduct a performance comparison between MARTIAN and its building blocks, i.e., LSTM and MLP, to show the effectiveness of the multi-view architecture in this problem domain. We also compare its predictive performance with some strong deep learning models, namely TabNet [102] and Gated Recurrent Unit (GRU) [103], which are shown to work well on the tabular data and time-series data, respectively. Please note that the input of time-series models at time-step t is a concatenation of E_t^c,L_t^c, and ACS_t^c. However, for the remaining models, the input is a concatenation of static features and all k steps of time-series inputs.

Table 4.2 compares MARTIAN with various baseline models for $n \in \{1, 2, 3\}$. We use two metrics to evaluate the predictive performance of forecasting models: (1) RMSE, which intuitively measures the average difference between each model's predictions and the actual number of eviction filings, and (2) Spearman correlation[8], which intuitively shows the extent to which the forecasted values preserve the actual orders of census tracts in terms of the number of eviction filing values. In this table, one row is considered for each ML model of interest and each column corresponds to the value of a performance metric for a specific value of n. Also, the best performance is shown in bold and the last row (Gain) shows the percentage of improvement that MARTIAN achieves over the best-performing baseline model.

[8]The value of Spearman ranges between -1 and 1. A higher Spearman shows a better performance of a forecasting model.

Model	n = 1		n = 2		n = 3		Avg. ($n \in \{1, 2, 3\}$)	
	RMSE	Spearman	RMSE	Spearman	RMSE	Spearman	RMSE	Spearman
Ridge	6.711	0.610	7.266	0.588	7.251	0.253	7.076	0.483
SVM	5.985	0.588	6.566	0.547	6.533	0.538	6.361	0.557
XGBoost	4.881	0.679	4.819	0.676	4.832	0.660	4.844	0.671
Random Forest	4.717	0.688	4.735	0.680	4.782	0.670	4.744	0.679
LightGBM	4.869	0.681	4.893	0.667	4.822	0.666	4.861	0.671
MLP	4.652	0.645	4.759	0.540	4.770	0.645	4.727	0.610
LSTM	4.585	0.639	4.717	0.639	4.753	0.631	4.685	0.636
GRU	4.590	0.649	4.686	0.648	4.755	0.631	4.677	0.642
TabNet	4.955	0.541	5.106	0.460	4.998	0.520	5.019	0.507
MARTIAN	**4.383**	**0.697**	**4.444**	**0.686**	**4.503**	**0.673**	**4.443**	**0.685**
Gain (%)	4.40%	1.30%	5.16%	0.88%	5.25%	0.44%	5.00%	0.88%

Table 4.2: Performance comparison of forecasting models.

According to the results, MARTIAN outperforms all baselines for all different values of n; in particular, on average, MARTIAN outperforms the best-performing baseline model by achieving 5.00% smaller RMSE and 0.88% higher Spearman, which shows its superiority against several strong ML models for this problem domain.

Additionally, MARTIAN achieves a Spearman value of 0.685 (on average), which shows that the ranking of census tracts in terms of the number of tenants at risk of formal eviction is preserved to high extent in MARTIAN's output.

Furthermore, MARTIAN outperforms both MLP and LSTM models, which form its building blocks; i.e., on average, it achieves 6.00% lower RMSE and 12.29% higher Spearman than MLP and improves the predictive performance of LSTM by 5.16% and 7.70% in terms of RMSE and Spearman, respectively. This could show the value of using multi-view architecture for incorporating data sources of various resolutions, rather than treating all inputs the same.

Moreover, comparing the performance of classical models, we see that decision-tree based models outperform the other ones (i.e., SVM and Ridge) significantly; i.e., on average, decision-tree based models achieve 28.31% and 29.42% better RMSE and Spearman, respectively. This could show that in case of any difficulty in using deep learning, decision-tree based ensemble models could be more appropriate ML choices for this task. Also, in spite of its high performance in several other domains, TabNet achieves the poorest performance among all our deep learning-based baselines and ensemble models. Therefore, this attention-based model does not seem to be an appropriate choice for this case.

Model	n = 1		n = 2		n = 3		Avg. ($n \in \{1, 2, 3\}$)	
	RMSE	Spearman	RMSE	Spearman	RMSE	Spearman	RMSE	Spearman
MARTIAN	**4.383**	**0.697**	**4.444**	**0.686**	**4.503**	**0.673**	**4.443**	**0.685**
MARTIAN-w/o-View1	5.887	-0.346	5.891	-0.385	5.862	-0.489	5.880	-0.406
MARTIAN-w/o-View2	4.731	0.623	4.748	0.650	4.878	0.574	4.785	0.615
MARTIAN-w/o-View3	4.615	0.689	4.673	0.676	4.686	0.572	4.658	0.645

Table 4.3: The results of MARTIAN's ablation study.

4.6.3 Ablation Study

We also conduct an ablation study to assess the impact of each view on the MARTIAN's predictive performance. To this end, we remove one view each time, train the new model, and then, evaluate its performance. Table 4.3 represents the results of our ablation study for $n \in \{1, 2, 3\}$. According to the results, removing view1 (i.e., features extracted from the time-series data of eviction filings) leads to a significant decrease in the predictive performance of MARTIAN; i.e., it results in 32.34% increase in RMSE and 159.27% decrease in Spearman (on average). In particular, we observe that MARTIAN-w/o-View1 cannot preserve the rank of census tracts with respect to the number of eviction filings as it has a negative spearman value. This makes sense because the time-series data of eviction filings is the only input data available at our forecasting spatial and temporal resolutions, and the other two data sources (i.e., labor statistics and ACS) are unavailable either at the census tract level or at the temporal resolution of one month. Therefore, two other data sources can only provide a big picture and alone are not enough for accurately forecasting the eviction crisis at high spatial and temporal resolutions.

Additionally, removing view2 (i.e., features extracted from the labor statistics data) results in a 7.69% increase in RMSE and a 10.21% drop in Spearman (on average). Thus, as expected, incorporating the monthly status of employment helps enhance the predictive performance of MARTIAN, even though it is not available for each census tract and it only reports the work status for "Dallas-Fort Worth-Arlington". Furthermore, excluding view3 (i.e., features extracted from the ACS data) leads to a 4.83% increase in RMSE and a 5.83% decrease in Spearman (on average). Therefore, although the ACS data reports the annual conditions of each census tract, its information on renter-inhabited housing units (and their householders) is still helpful for predicting the number of eviction filings at the census tract level for each month. In conclusion, as a result of this ablation study,

Model	n = 1		n = 2		n = 3		Avg. ($n \in \{1, 2, 3\}$)	
	RMSE	Spearman	RMSE	Spearman	RMSE	Spearman	RMSE	Spearman
Ridge	6.620	0.596	6.528	0.582	6.725	0.561	6.624	0.579
SVM	6.298	0.618	6.374	0.588	6.467	0.564	6.379	0.590
XGBoost	5.268	0.667	5.238	0.655	5.276	0.632	5.260	0.651
Random Forest	5.126	0.675	5.062	0.662	5.068	0.653	5.085	0.663
LightGBM	5.145	0.653	5.251	0.634	5.180	0.655	5.192	0.647
MLP	4.941	0.639	4.944	0.640	5.013	0.629	4.966	0.636
LSTM	4.998	0.619	4.978	0.612	5.032	0.601	5.002	0.610
GRU	4.994	0.625	4.948	0.620	5.034	0.605	4.992	0.616
TabNet	5.371	0.450	5.541	0.390	5.466	0.381	5.459	0.407
MARTIAN	**4.827**	**0.698**	**4.755**	**0.688**	**4.823**	**0.680**	**4.801**	**0.688**
Gain (%)	2.30%	3.40%	3.82%	3.92%	3.79%	3.81%	3.32%	3.77%

Table 4.4: Performance comparison of forecasting models in the cross-region test.

we find that both labor statistics and ACS data are useful auxiliary input signals for our forecasting task.

4.6.4 Cross-Region Test

In all our previous experiments, we trained forecasting models on the training portion of the Dallas data, and then, evaluated their performance on the testing portion of the same data. We now conduct a cross-region test, in which the training and testing datasets are created from the data of two disjoint sets of census tracts. This helps us evaluate if MARTIAN's superior predictive performance is generalizable to unseen regions (whose data has not been seen by the model in the training phase). To this end, we take the following steps: (1) we create two disjoint sets of census tracts (with almost equal size) through random sampling such that the statistics (minimum, maximum, median, and average) of the total number of eviction filings for these two sets look similar, (2) we train the forecasting models on the training portion of the first set, and (3) we assess the performance of forecasting models on the testing portion of the second set. Please note that we still use the walk-forward testing approach and the time frame of training and test sets is the same as before.

Table 4.4 shows the results of our cross-region test. According to the results, MARTIAN outperforms all baseline models for different values of n; in particular, on average, it achieves 3.32% lower RMSE and 3.77% higher Spearman than the best-performing baseline model. This shows that MARTIAN's superior predictive

performance is generalizable to various unseen regions.

4.7 Real-World Use Case

Our tool could serve as an AI assistant to (1) shed light on the number of tenants at risk of getting formally evicted in the future; e.g., the output of MARTIAN can be used to generate a heatmap of the forecasted number of tenants at-risk of eviction for each month in the future (similar to Figure 4.2), and (2) make a more well-informed resource allocation plan to mitigate evictions in a more efficient and effective manner. In particular, we contacted officials at Texas Housers (i.e., Texas Low Income Housing Information Service)[9], which is an organization aiming at mitigating housing problems in Texas. Ben Martin, who is an official at Texas Housers and is working on the eviction and foreclosure data, stated that:

> "Knowing where evictions are being filed helps advocates, administrators, elected officials, and legal aid to identify where they need to direct their efforts, funds, and other resources in order to keep renters housed"

In particular, he elaborated on the potential impacts of such forecasting tools in the real world as follows:

> "The number of eviction cases filed or of a certain outcome, might, for example, be used as a baseline for setting program funding levels. If a somewhat accurate tool could be developed, it would be incredibly useful for advocacy with the legislators, elected officials, and agencies responsible for eviction court and eviction diversion"

we also conduct a synthetic simulation to get a better estimate of the value of our predictive approaches for resource allocation. In fact, we assume that each at-risk tenant only gets $100, and the funding is allocated exactly based on the predicted number of tenants at-risk of eviction, and then, by comparing the results with the ground-truth numbers of eviction filing, we estimate the amount of unused funding and shortage resulting from each approach. In particular, we compare

[9] `https://texashousers.org`

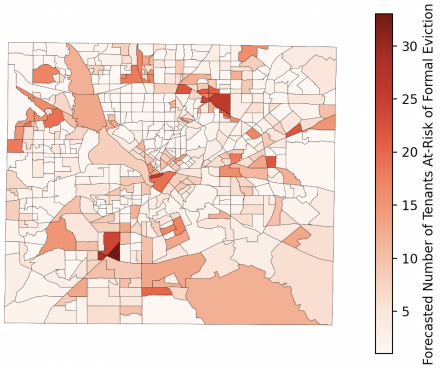


Figure 4.2: MARTIAN's forecasts about the number of tenants at-risk of formal eviction at various census tracts within Dallas County, TX in December 2021.

MARTIAN and LSTM for the task of first-step forecasting, and according to the empirical results, it seems that using MARTIAN results in about 4.4% and 5.6% decrease in unused funding in some regions, and shortage in funding in some other regions (compared to LSTM), respectively.

Therefore, all these pieces of evidence show the value of an accurate eviction forecasting tool and the extent to which it could help policymakers enhance eviction diversion/prevention programs in the field.

4.8 Summary

This chapter developed a neural network model, named as MARTIAN, that leverages data sources of various resolutions and forecasts the number of tenants at-risk of getting formally evicted at the census tract level n months into the future in a fully-supervised manner. The results of our empirical evaluation show that MARTIAN outperforms various baseline models in terms of RMSE and Spearman in all considered situations. Additionally, the results of our cross-region test show that MARTIAN's superior predictive performance is generalizable to unseen census tracts. MARTIAN could help policymakers direct funding and other resources in a more efficient manner and enhance the existing eviction prevention/diversion programs by providing data-driven insights on the future condition of each census tract in terms of eviction filings.

Chapter 5
AI for Social Welfare of Housing-Insecure Low-Income Americans: Eviction Filing Prediction with Coarse-Grained Ground-Truth Labels

Similar to the previous chapter, this chapter develops predictive ML models to forecast the number of tenants at-risk of eviction at a high resolution. However, in this chapter, we assume that ground-truth eviction filing data is only available at a low spatial resolution, rather than high resolution [5]. In the following sections, we describe the problem domain, related work, our solution, experimental results, and some real-world implications.

5.1 Introduction

To help mitigate the eviction crisis, Chapter 4 develops predictive ML models to forecast the number of tenants at-risk of formal eviction. Relying on the fully-supervised learning approach, it mainly assumes that ground-truth labels are available at the spatial resolution of interest. However, this assumption is somewhat strong because, for many regions, individual-level eviction data is not readily available and the spatial resolution of available data might be much lower

than the resolution of interest, which in turn, adversely affects the performance of some commonly-used fully-supervised learning methods (such as the MSE loss function alone) in this situation.

To address this challenge, recent research [23] proposed a *coarsely-supervised training approach* for a regression task; i.e., it proposes a loss function to ensure that the predictive model returns similar values for similar data points, while the average of predictions at a low resolution is close to ground-truth. While being effective for the task of vegetation monitoring, according to our empirical evaluation, this solution does not seem to perform well in our problem domain. Consequently, inspired by this method [23], we propose a loss function that leverages low-resolution ground-truth eviction data as well as sociological insights to facilitate the process of training neural networks in the face of a lack of high-resolution ground-truth labels. To be more specific, we use a proxy factor that tends to be positively associated with eviction and is available at a high resolution. Then, our loss function tries to ensure that model's predictions can preserve the ranking of data points with respect to that proxy factor, while the model's of predictions is close to the ground truth at a low spatial resolution (similar to [23]).

We conduct various experiments to analyze the effectiveness of our solution under various conditions. According to the results, leveraging a proxy factor in the loss function results in considerable improvement in predictive accuracy. Furthermore, we analyzed the link between our loss function's effectiveness and the level of association between the proxy factor and the target variable. As a result, we find that our proxy factor's association with eviction indeed plays a key role in the effectiveness of our solution. Then, to get a better understanding of the real-world implications of this approach, we also conduct a synthetic simulation to compare the value of incorporating our predictive approach for allocating funding, and the results suggest that our approach could potentially have significant positive impacts on enhancing resource allocation and reducing the amount of unused funding (compared to some considered baselines).

5.2 Related Work

There has been extensive research on mitigating the eviction crisis or addressing the challenge of a lack of high-resolution labels. This section surveys several recent

studies under these two categories.

Research on the Eviction Crisis. As mentioned in the previous chapter, many of the prior studies on the eviction crisis come from the social science discipline, where scholars studied the association of various factors (such as job loss, crime rate, etc.) with eviction [88] and the consequences of eviction [92]. While highly insightful, those works mostly rely on statistical analysis and prevalence studies and did not focus on forecasting future conditions in terms of eviction. Additionally, there has been a couple of studies that relied on ML approaches to help mitigate the eviction crisis [80, 104], however, they assume that enough ground-truth labels are available at the spatial resolution of interest, which limits their usability in this problem domain.

Research on ML under Lack of High-Resolution Labels. In the ML literature, there are numerous studies on facilitating the training of ML models with limited/no ground-truth labels. Some research proposed to train neural networks on a proxy factor associated with the target variable, and then, fine-tune their weights using a limited number of ground-truth labels [105, 106]. While effective, they assume that a small number of ground-truth labels is available at the instance level, which does not align with the assumptions mentioned in this chapter. Additionally, there has been research on multiple-instance learning [107] and weakly-supervised learning [6, 108], however, they mostly focused on the classification task. Recently, Fan et al. [23] proposed a coarsely-supervised training approach, which is very close to our solution; i.e., they suggested a new loss function, which penalizes the prediction of too different values for similar inputs (through a smoothness loss term), while trying the preserve the model's predictions close to the ground-truth at a low resolution. However, our empirical results show that it does not seem to work well in our problem domain. Inspired by this approach, we propose a new loss function that relies on a proxy factor, rather than input similarity, to capture differences among training data points.

5.3 Datasets

Similar to the previous chapter, we relied on three datasets.

(1) Eviction Filing Data. We use the eviction filing data of Dallas County, TX (from 2021 to 2022), released by CPAL at `https://northtexasevictions.org/`. It includes the monthly number of eviction filing for each census tract. Please note that, during training, we assume that, for each month, only the total number of eviction filings for the entire county is available, and *we only use the high-resolution ground-truth number of eviction filings to evaluate the accuracy of different neural network models in the testing phase.*

(2) American Community Survey (ACS). As mentioned before, this data[1] consists of various pieces of information about the characteristics of renter-occupied housing units and the work status. Its temporal resolution is one year, and in this study, we rely on the data of 2020.

(3) Labor Statistics. This data[2] includes monthly statistics on the unemployment rate and employment rate in various fields. Due to the strong association between work status and eviction [88], this source of data tends to be a useful auxiliary signal for forecasting the number of eviction filings [80].

5.4 The Proposed Methodology

In this section, we propose a new loss function that can be used for training various neural network models. This loss function incorporates low-resolution eviction data as well as a high-resolution proxy factor to facilitate the process of training in the face of a lack of high-resolution ground-truth labels. In this section, we, first, define some formal notations, and then, describe our loss function.

Formal Notation. For the task of 1^{st}-step forecasting, we show each training data point as follows:

$$(\{x^{Eviction}_{Dallas,\{t-k,...,t-1\}}, x^{Labor}_{Dallas,\{t-k,...,t-1\}}, ..., x^{ACS}_{c_i,t-1}\}, y_{c_i,t})$$

In this formula, $x^{Eviction}_{Dallas,\{t-k,...,t-1\}}$ refers to a vector of length k that represents the average number of eviction filings in Dallas County over the preceding k months, and $x^{Labor}_{Dallas,\{t-k,...,t-1\}}$ refers to a vector of length k that represents the labor statistics

[1] `https://www.census.gov/programs-surveys/acs/data.html`

[2] `https://www.bls.gov/eag/eag.tx.htm#eag_tx.f.2`

over the past k months, $x^{ACS}_{c_i,t-1}$ denotes the corresponding ACS factors, and $y_{c_i,t}$ refers to the total number of eviction filing in census tract i at month t. Please note that, as mentioned before, *we assume that, during training, $y_{c_i,t}$ is unknown, and we only know the total (or average) number of eviction filings in Dallas County at month t.*

The Proposed Loss Function. Inspired by [23], our proposed loss function is a weighted sum of two terms as follows: $L_{coarse} + \alpha \times L_{pairewise-ranking}$. Similar to [23], the first term (L_{coarse}) ensures that, at each point in time, a model's prediction is close to the ground-truth value at a low spatial resolution (i.e., at the county level) and is defined as the mean squared error between the average of model's prediction for census tracts within Dallas County and the corresponding ground-truth. However, this term is not enough for capturing differences among various census tracts of a county; i.e., a model can output the same prediction for all census tracts, while ensuring that their average is close to that county's average number of eviction filings, and this is not a desirable result.

To circumvent the aforementioned challenge, we introduced the second term ($L_{pairewise-ranking}$), which aims to ensure that a model's prediction preserves the ranking of census tracts in terms of eviction. To this end, we take the following major steps. First, similar to the previous chapter, we review social science literature to find what factors tend to be highly associated with eviction, and then, try to look them up in ACS data. If the exact factor is not found, we choose a semantically close factor instead. Then, one factor is selected as the proxy factor, and in this chapter, we choose the number of renter-occupied housing units, for which the housing cost is more than 30% of householder's income because prior work showed that more than 70% of low-income renters devote more than half of their income on housing expenses and eviction tends to be prevalent among low-income population, and so, it looks close to sociological insights [6, 20, 21, 109]. Next, we split the range of its value into n bins[3]. Then, $L_{pairewise-ranking}$ intuitively aims to ensure that, for each census tract, the average of predicted values over one year is positively ranked with respect to the value of the proxy factor. In fact, it is defined as the aggregation of pair-wise ranking loss [110, 111] between the data points of two consecutive bins.

[3] n is a hyper-parameter, which is set through hyper-parameter tuning.

5.5 Experimental Evaluation

This section provides the results of various experiments that we conduct to analyze the effectiveness of our loss function under different conditions. First, we train multiple neural network models with various loss functions and compare our loss function with a number of baselines. Then, we analyze the impact of the choice of proxy variables and its association with eviction on the effectiveness of our solution. We also conduct a simulation to provide insights on the impact of predictive modeling and the achieved improvements on enhancing funding allocation.

5.5.1 Set-Up

Following prior research [80], we use the walk-forward testing approach [74] (the window size is set to 1; i.e., one month). Additionally, we use the Adam optimizer [98] with a learning rate of 10^{-4}, and the early stopping approach [99]. Finally, n (number of bins for splitting the proxy variable), k (length of time-series historical data), and the maximum number of epochs are set to 10, 6, and 1000, respectively. All experiments are run 3 times and the average is reported in the following subsections.

5.5.2 Comparison with Baseline Models

In this section, we compare our loss function with the following ones: (1) Coarse Loss, which only uses the low-resolution labels during optimization, (2) Pairwise ranking loss, which is the second term of our proposed loss function, and (3) Coarsely-Supervised approach [23], which assign large loss values if model's predictions for similar inputs differ a lot. We adopted the original method to our time-series situation as follows: At each point in time, the similarity of two census tracts is defined as the euclidean distance between their input features, and the similarity of two data points at different time frames is considered to be 0.

These loss functions can be used for training different types of neural network models, and in this work, we trained the following models: (1) Multi-Layer Perceptron (MLP), (2) Long Short-Term Memory (LSTM) [34], (3) Gated Recurrent Units (GRU) [103], and (4) MARTIAN [80].

Step	Loss Function	MLP		LSTM		GRU		MARTIAN	
		RMSE	spearman	RMSE	spearman	RMSE	spearman	RMSE	spearman
1	Coarse Loss	8.480	0.151	8.466	0.121	8.477	0.091	8.459	0.196
	Pairwise Ranking	9.150	0.520	9.300	0.523	9.191	0.542	8.986	0.489
	Coarsely Supervised	8.471	0.050	8.464	0.163	8.477	-0.023	8.458	0.263
	Our proposal	8.221	0.452	8.145	0.533	8.138	0.511	8.100	0.527
2	Coarse Loss	8.475	-0.124	8.470	0.158	8.464	0.000	8.452	0.249
	Pairwise Ranking	9.098	0.505	9.305	0.529	9.286	0.530	8.987	0.490
	Coarsely Supervised	8.480	-0.235	8.459	0.327	8.443	0.415	8.449	0.245
	Our proposal	8.234	0.542	8.163	0.531	8.231	0.531	8.167	0.514
3	Coarse Loss	8.465	0.113	8.471	-0.051	8.459	0.041	8.448	0.321
	Pairwise Ranking	9.086	0.512	9.319	0.516	9.237	0.543	8.987	0.491
	Coarsely Supervised	8.453	0.140	8.459	0.279	8.444	0.341	8.445	0.427
	Our proposal	8.290	0.380	8.163	0.528	8.169	0.534	8.133	0.517

Table 5.1: The accuracy of various neural networks with different choices of the loss function on the n^{th}-step forecasting task.

Table 5.1 shows the performance of various neural networks with different loss functions on the n^{th}-step forecasting task. According to the results, our proposed solution outperforms the coarsely supervised approach [23] in all considered situations, and on average, it achieves 3.2% and 155.2% better RMSE and spearman than the coarsely-supervised approach, respectively. Furthermore, comparing coarsely supervised training [23] and coarse loss, we see that, in many cases, the smoothness term in [23] helps improve the predictive performance, however, it is less effective than the pairwise ranking terms proposed in this study. Additionally, the smoothness term seems to be effective for capturing the ranking (to some extent), however, it does not seem to help much in improving RMSE. In contrast, our proposed ranking loss term helps improve the predictive performance of the underlying model in terms of both RMSE and spearman.

Additionally, the results suggest that using either coarse loss or pairwise ranking is not enough for accurate prediction. In fact, while pairwise ranking loss is helpful for preserving the ranking of predictions in terms of the number of eviction filings, as expected, the predicted values are not necessarily close to the actual values, and removing the coarse loss results in 12% increase in RMSE, on average. Also, using coarse loss is not enough for preserving the ranking of data points in terms of the number of eviction filings, and removing the pairwise ranking loss terms results in 79.3% decrease and 3.4% increase in spearman and RMSE, respectively (on average), and the spearman value becomes 0.1, on average. Therefore, our solution uses both terms to be able to improve both performance metrics.

5.5.3 Impact of the Choice of Proxy Variable

While the previous experiments confirm the effectiveness of our solution, compared to some baselines, we hypothesize that the effectiveness of our solution is closely associated with the proxy factor's correlation with the number of evictions in our dataset. To check this hypothesis, first, we select a couple of ACS factors that have different levels of association with the number of eviction filings. Then, we train a separate MARTIAN using each of these ACS factors and evaluate its performance on the task of first-step prediction.

Figure 5.1 shows changes in the values of RMSE and spearman with different choices of proxy factors. In this figure, the y-axes show the RMSE and spearman values and the x-axis represents a correlation metric between one proxy factor and the average number of eviction filings according to our dataset. As expected, the correlation between the proxy variable and eviction plays a key role in the effectiveness of our proposed solution.

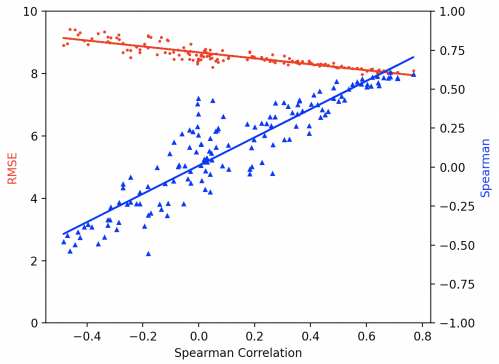


Figure 5.1: Effectiveness of our solution with different proxy factors with various levels of association with the number of eviction filings.

5.6 Real-World Use Case

As described in the previous chapter, such predictive models could provide insights to policymakers regarding future conditions in terms of the potential number of tenants at-risk of eviction and help them in their efforts at mitigating the eviction crisis across the United States. In particular, efficient resource allocation is a challenging issue in the field; e.g., a large difference has been observed in the use of

allocated function, and some regions sent back millions of dollars of rental assistance funding to the government [80]. As a result, it is important to enhance the existing programs [84], and an accurate predictive tool could assist in allocating funding in a more efficient/effective way [80]

In this section, we also conduct a synthetic simulation to get a better estimate of the value of our predictive approaches for resource allocation. In fact, we assume that each at-risk tenant only gets $100, and the funding is allocated exactly based on the predicted number of tenants at-risk of eviction, and then, by comparing the results with the ground-truth values, we estimate the amount of unused funding and shortage resulting from each approach. In particular, we use the output of MARTIAN for the first-step prediction task with different loss functions in this analysis.

Table 5.2 shows the amount of demanded funding, unused funding (expected to be returned from some census tracts), and shortage in funding (in some other census tracts). According to the results, on average, the model trained on the coarse loss demands $356,800, but about $168,100 remains unused, while $194,300 more funding is needed in other regions. However, if we repeat the same process using the model trained on our proposed loss function, it demands $307,600, but $123,000 remains unused, and $198,400 more money is needed from other regions (on average). As a result, we see that our proposed solutions could help reduce the amount of unused funding by 26.7%, and fulfill the demands of other regions comparably, although it demands 13.7% less funding (in total), compared to the coarse loss case. Additionally, the empirical results show that the model trained on coarse loss works similar to the model trained on coarsely-supervised loss in this case. Furthermore, as expected, incorporating the pair-wise ranking alone is not enough to accurately estimate the total amount of required funding.

Loss Function	Demanded Funding	Funding Shortage	Unused Funding
Coarse Loss	$356,800	$194,300	$168,100
Pairwise Ranking	$142,100	$284,400	$43,500
Coarsely Supervised	$355,000	$194,600	$166,700
Our Proposal	$307,600	$198,400	$123,000

Table 5.2: Real-world impact of various loss functions for resource allocation.

In conclusion, as a result of these analyses, we hypothesize that using our

proposed solution at scale could lead to a major improvement in allocation funding compared to using the coarse loss (or the coarsely-supervised method [23]).

5.7 Summary

This chapter proposes a new loss function to facilitate the training of neural network models under a lack of ground-truth labels at the spatial resolution of interest. In particular, it leverages sociological insights as well as low-resolution labels to accurately forecast the number of eviction filings at a high resolution under a lack of access to high-resolution ground-truth labels. Our empirical evaluation shows that it highly outperforms a recent approach in terms of RMSE and Spearman. We also conduct a simulation to assess the value of using such predictive models for funding allocation, and we find that using the proposed loss function could be a better choice to enhance resource allocation programs, compared to some considered baseline methods.

Chapter 6
AI for Social Welfare of Housing-Insecure Low-Income Americans: Eviction Filing Hotspot Detection with No Ground-Truth Labels

Previous chapters focused on developing predictive ML models when either low-resolution or high-resolution labels are available for training ML models of interest. However, this chapter focuses on the eviction crisis, and aims to detect eviction filing hotspots from publicly available data under a lack of ground-truth labels [6]. In the following sections, we describe the problem domain, related work, our solution, experimental results, and real-world application.

6.1 Introduction

Numerous low-income renting families across the USA are at a high risk of eviction, mainly due to a shortage of federal housing assistance, and an ever-increasing gap between income growth and increases in housing cost, e.g., about 70% of the low-income renters devote most of their income towards housing expenses [20, 21]. Further, eviction is an important cause of several societal problems, such as homelessness [85, 86], and has long-lasting negative effects on individual's health,

and housing stability [20, 22, 81, 82]. As a result, mitigating the eviction crisis is of the utmost importance in order to enhance the well-being of this community.

To tackle this crisis, NGOs and policymakers have been implementing multiple programs at the pre-filing and post-filing stages [112–114]; especially, they assign several types of resources to increase housing stability and affordability. Efficient resource allocation for these eviction prevention programs is possible with a data-driven understanding of eviction filing hotspots across the USA. Unfortunately, there is no national eviction database [115] and some state/local policies and resource limitations restrict access to ground-truth eviction filing records for many regions of USA. For example, in Illinois, bulk data retrieval is not allowed or in California, tenants may block public access to their eviction records [97]. Furthermore, the high cost of data collection in some other regions makes it infeasible to collect eviction filing records at scale, e.g., obtaining eviction filing data from some courts requires in-person data collection [97]. These obstacles limit our understanding of eviction filing hotspots in those regions [97], which in turn, calls for a need for some solutions to fill this gap and help policymakers more effectively/efficiently implement eviction diversion programs under a lack of access to court records.

To this end, this chapter proposes WARNER (**W**eakly-supervised **A**id to **R**elieve **N**ationwide **E**viction **R**ate), a weakly-supervised ML model that leverages publicly available satellite imagery as well as sociological insights (instead of ground-truth labels) to predict eviction filing hotspots across the USA. In fact, this chapter makes the following contributions: (1) to account for the lack of sufficient labeled training data in this domain, it proposes a label generation approach that leverages the findings of past literature in sociology to produce high-quality labels for a subset of unlabeled training data, (2) it develops a neural network model to predict eviction filing hotspots from satellite imagery of different shapes, and (3) it does several experiments to assess the accuracy of WARNER using a real-world dataset with eviction filing records in Dallas County, TX.

Our empirical evaluation shows the high quality of the labels generated by our proposed label generation approach. Furthermore, it shows that WARNER outperforms multiple strong baseline models by obtaining about 36.0% and 1.4% higher F1 and AUC (respectively), which illustrates its suitability for this domain. Additionally, the superior accuracy of WARNER can be generalized to different counties across the USA. This work is conducted in collaboration with CPAL.

6.2 Related Work

This section surveys past literature in the areas of sociology and Machine Learning.

Sociological Research Prior work in sociology mostly focuses on understanding the factors associated with eviction using statistical and descriptive analysis. In fact, prior work has studied the association between the risk of eviction and various individual and neighborhood-level characteristics. e.g., they found that job loss and crime rates in a neighborhood tend to increase the risk of eviction [88]. Further, prior research found that low-income single mothers who have young children tend to be at a high risk of getting evicted [116]. Additionally, according to their findings, eviction could lead to long-lasting health problems (e.g., depression) [20]. Even though this line of work gleaned unique insights about the eviction crisis, they did not address the problem of predicting eviction (or eviction filing) hotspots across the USA (which is the focus of our work).

Some other prior work focuses on finding eviction hotspots in certain geographic regions by counting the total number of eviction filings in their sub-regions. As a result, they find that a large number of evictions in a region can be attributed to a small number of sub-regions [82, 117]. However, these works require the actual number of evictions (or eviction filing records), which is inaccessible (or highly expensive to obtain) for many regions due to restrictive state/local policies and infrastructure limitations [97]. Therefore, their methodology is not generalizable to all regions within USA. In contrast, this chapter proposes a highly generalizable ML-based framework that relies on satellite imagery and sociological insights (rather than the actual number of eviction filings) to predict eviction filing hotspots within US counties in the absence of court records.

Machine Learning Research There has been a large number of research on applying ML techniques to tackle societal problems. One line of research developed predictive ML models using a tabular dataset consisting of several factors with potential impacts on the dependent variable [93, 94]. For example, Ye et al. [93] relied on classical ML models to predict the risk of landlord harassment using a tabular dataset. However, these models have limited real-world usability, as their predictive performance is highly dependent on data sources that are either (1) unavailable for many regions across the USA, or (2) highly expensive to obtain

as they need to be gathered by conducting surveys. Additionally, earlier in this dissertation, we developed neural network models that forecast the number of eviction filings for each census tract. However, we assumed that the historical eviction filing data is available for the target region, but this assumption does not necessarily hold at the national scale. In contrast, in this chapter, we rely on publicly available datasets that cover all census tracts across the USA (namely, satellite imagery and the American Community Survey data[1]).

Another line of research takes advantage of imagery data and variants of Convolutional Neural Network (CNN) models [118] to predict factors related to poverty and human development. In particular, some prior work focused on predicting poverty from satellite imagery in the face of sparsely labeled data [105, 106], and to tackle this issue, they proposed to incorporate night-time light intensity as a proxy for poverty during training. However, a subsequent study [119] showed that this methodology does not necessarily generalize to predicting some other human development factors (such as access to water and average child weight-to-height percentile). Additionally, some studies used computer vision approaches (such as object detection techniques [120, 121], panoptic image segmentation [122], and CNN-based neural networks [123, 124]) for predicting poverty and/or other development indicators from imagery data. However, relying on a supervised learning paradigm, all these studies trained their models on a dataset consisting of ground-truth labels. In contrast, this chapter proposes a framework for predicting eviction filing hotspots without access to ground-truth labels; instead, it addresses the lack of labeled training data by leveraging insights from prior work in sociology to develop a weak supervision approach to generating labels.

6.3 A Problem Statement

This section provides a formal definition of the problem of identifying eviction filing hotspots in US counties. Intuitively, **eviction filing hotspots** of a county c over a period of m years refer to the census tracts (in that county c) which "*consistently have high contributions*" to the total eviction filings in c during a period of m years.

More formally, we define **top-k% eviction filing hotspots** of a county c over a period of m years as follows. Suppose that c has n census tracts, and E_t^i denotes the

[1] `https://www.census.gov/programs-surveys/acs/data.html`

total number of eviction filings in the i^{th} census tract of c in year t ($t \in \{1, ..., m\}$) after sorting census tracts of c in the descending order of their number of eviction filings in year t. Additionally, let S_t^d refer to the largest set of census tracts (in descending order) whose combined number of eviction filings is less than or equal to $d\%$ of the total number of filings in c in year t (i.e., $\sum_{tract_i \in S_t^d} E_t^i \leq \frac{d}{100} \times \sum_{i=1}^{n} E_t^i$). Please note that we add census tracts to S_t^d in decreasing order of the number of eviction filings. Then, the *top-k% eviction filing hotspots of c over a period of m years* are defined as the set $\cap_{t=1}^{m} S_t^d$ such that $|\cap_{t=1}^{m} S_t^d| \approx \lceil \frac{k \times n}{100} \rceil$. In this definition, k and m are considered to be fixed (defined by stakeholders) and d is chosen to be the largest number such that $|\cap_{t=1}^{m} S_t^d| \leq \lceil \frac{k \times n}{100} \rceil$. Note that these conditions can be satisfied with fractional values of d and k.

Finally, we formulate the problem of identifying top-k% eviction filing hotspots as a *binary classification problem*, in which the ultimate goal is to predict if a census tract belongs to the top-k% eviction filing hotspots of its county (i.e., positive label) or not (i.e., negative label). This chapter builds an ML model that takes the satellite images of a census tract ($\{x_1^{tract}, x_2^{tract}, ..., x_m^{tract}\}$) and its county ($\{x_1^{county}, x_2^{county}, ..., x_m^{county}\}$) as input and outputs a prediction for the binary variable of interest. To assess the effectiveness of the proposed model for this problem domain, we experiment with different values of k in Section 6.6.

6.4 Datasets

In this study, we use three datasets: (1) American Community Survey data, (2) Satellite imagery, and (3) Eviction filing records.

American Community Survey (ACS) As mentioned before, ACS data contains various pieces of information on demographic characteristics, housing characteristics, work status, and poverty status in the past 12 months. This dataset is published by the U.S. Census Bureau annually, but with a delay of about two years. We use ACS 5-Year Experimental Estimates in this work as it provides annual statistics for all census tracts in the U.S. *Note that we only use the ACS data to generate weakly supervised labels for our satellite imagery training datasets, which we describe next.*

Satellite Imagery We use Sentinel imagery[2], which provides a bird's eye view

[2] `https://sentinel.esa.int/web/sentinel/missions/sentinel-2`

of the environment with spatial and temporal resolutions of 10 meters and 10 days, respectively. For each census tract, we crawl one image corresponding to the bounding box of that tract (i.e., the minimum rectangle surrounding the polygon of that census tract). Further, for each image, we generate a mask matrix to be able to distinguish the pixels that fall inside that census tract (i.e., valid pixels) from the other ones (i.e., invalid pixels).

Eviction Filing Records This dataset consists of individual eviction cases filed across Dallas County, TX since 2017 and we got access to this data through our collaboration with CPAL. Each record contains information about the eviction filing time, tenant's address (i.e., latitude and longitude), names of both parties (i.e., landlord and tenant), etc. *Please note that while this dataset on eviction filing records is available for Dallas County (through our collaboration with CPAL), getting similar datasets from other U.S. counties is very challenging, if not impossible. Since we want a generalizable ML model that can predict eviction filing hotspots across all US counties (not just Dallas County), we do not use this dataset to train our ML model. Instead, we only use this source of data for evaluating the performance of WARNER.*

Data Preparation We now explain our data preparation process. For each census tract (and county), we consider the median of the three least cloudy satellite images collected from the beginning of June to the end of July of a year[3] as the satellite image of that census tract (and county) in that year. Then, we convert the value of each pixel into the range of $[0, 1]$. Additionally, to prepare the eviction filing records of Dallas county (for which we have ground-truth labels), we take three main steps. First, following [97, 125], we exclude the eviction cases filed against business defendants and remove duplicates. Then, for each census tract, we calculate the number of eviction records in each year. Finally, the data is split into the ratio of 60:20:20 while keeping the class distribution among train, validation, and test sets.

[3]We use the satellite images taken in summer because, over that period of time, the climate condition seems to be suitable across the USA for taking clear images.

6.5 The Proposed Framework: WARNER

We propose a weakly-supervised framework to address the problem of predicting eviction filing hotspots from satellite imagery in the face of a lack of ground-truth eviction filing data. Figure 6.1 illustrates the architecture of WARNER, which is composed of two components: (1) a *label generation model* that generates probabilistic labels for a subset of an unlabeled satellite imagery dataset by leveraging insights from prior work in sociology as well as the ACS data, and (2) a *hotspot prediction model* that predicts eviction filing hotspots from satellite imagery (along with the generated labels) using a neural network model. In the following subsections, we elaborate on the architecture of each component.

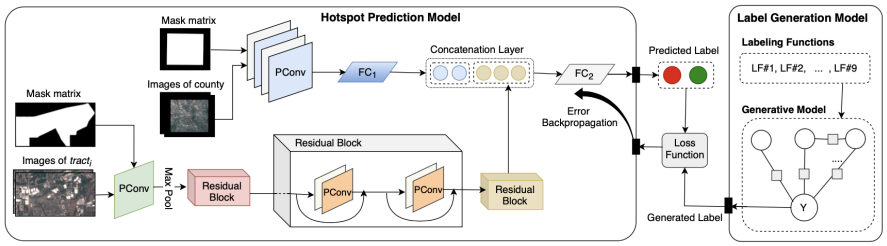


Figure 6.1: The architecture of WARNER.

6.5.1 The Label Generation Model

As the first step toward predicting eviction filing hotspots, we build an ML model that uses sociological insights along with the ACS data (as input) to generate labels for our satellite imagery training dataset. In fact, the following steps are taken: (i) we survey past literature in sociology to find several factors that are highly associated with eviction. (ii) Then, we define one labeling function (LF)[4] for each associated factor; i.e., a labeling function labels each data point based on the value of the underlying associated factor. (iii) Finally, we use a weak supervision framework via Snorkel [108] to combine the results of various labeling functions; i.e., since each data point might be labeled by several labeling functions, we use

[4]A labeling function is a piece of code that takes a data point (i.e., census tract) as input and assigns a label (positive, negative, or abstain for binary classification) using some rules (or heuristics, etc.).

Snorkel to convert that set of *potentially noisy* labels into one probabilistic label. The following paragraphs provide further details regarding each step.

Sociological Insights. To mitigate the eviction crisis, sociologists and social work scientists have been studying various aspects of the eviction crisis and housing instability. As a result, they have discovered several data-driven insights; for example, they found a high level of association between the risk of eviction and some demographic and financial characteristics of renters (and neighborhoods) [88, 90, 95]. However, these studies rely on datasets collected through conducting in-person surveys from a relatively small population, and thus, their datasets and studied features are not available as-is for all U.S. census tracts. To tackle this challenge, we use the ACS dataset, which consists of various demographic, financial, and housing characteristics at the census tract level (across the entire U.S.). Next, we review prior work in sociology to find a set of factors associated with eviction and housing instability [21, 88, 90, 95, 96]. For each associated factor, we try to locate that factor among the set of features present in the ACS dataset. If an associated factor is not found as-is in the ACS dataset, an ACS feature that is semantically close to that factor is selected, instead. Note that in spite of their association with eviction, some neighborhood-level characteristics (such as past eviction rate) [88] and social network properties (such as network disadvantage) [88] are not considered in this study because they are not gathered in the nationwide ACS dataset. As a result of taking these steps, we successfully locate nine associated factors (as reported in prior sociology literature) in the ACS dataset (similar to [125]), and these nine factors form the basis of our label generation model.

Table 6.1 provides the definition of these nine associated factors that were located in the ACS dataset. Each of the selected factors is shown to have some sort of association with eviction. For example, while job loss, and hence, zero income (LF#2) tend to increase the risk of eviction [88, 90], being employed (LF#9) has shown to be a protective factor[5] for housing instability [95]. Additionally, most low-income renting families reportedly spend a considerable amount of their income on housing expenses; in fact, about 70% of them devote most of their earnings on housing expenses [21]. Accordingly, we include LF#3 in our feature set. Furthermore, individuals with educational attainment of less than high school

[5]Protective (risk) factors refer to factors that are associated with a lower (higher) chance of a negative outcome.

(LF#4) tend to be at higher risk of eviction [90], whereas having higher level of education (LFs #5, #6, and #7) tends to be associated with a lower likelihood of housing instability [96]. Intuitively, census tracts with higher numbers of renter-inhabited housing units (LF#1) tend to have high contributions to the county's total evictions. Finally, recipients of public assistance (LF#8) are found to be at lower risk of housing instability [95, 96]. *Please note that due to the potential ethical implications of labeling data points only based on some protected characteristics (such as gender, race, and age), we did not utilize such factors for defining labeling functions.* Next, we describe how we use this set of associated factors to generate probabilistic labels for an unlabeled dataset.

Table 6.1: The definition of ACS factors underlying our labeling functions.

LF#	Explanation of the Underlying ACS Feature	Polarity
1	# of housing units occupied by renters	Positive
2	# of renter-occupied units whose householder has zero or negative earnings in the previous 12 months	Positive
3	# of renter-occupied units, with monthly housing costs $\geq$ (0.3×income)	Positive
4	# of renter-occupied units whose householder's level of education is less than high school	Positive
5	% of renter-occupied units whose householder is a high school graduate (or equivalent)	Negative
6	% of renter-occupied units whose householder has a college or associate's degree	Negative
7	% of renter-occupied units whose householder has at least a bachelor degree	Negative
8	% of families below poverty line who get paid SSI or cash public assistance income	Negative
9	% of full-time workers with some earnings	Negative

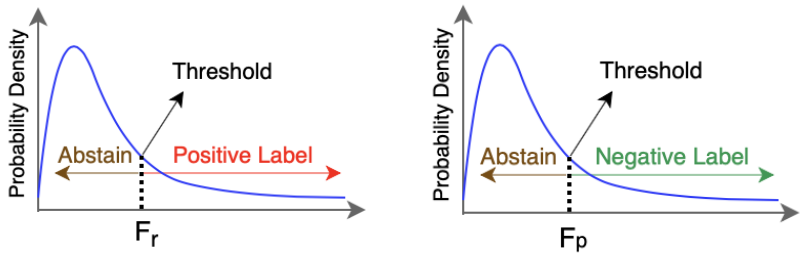
Design of Labeling Functions. Although prior work distinguishes between protective and risk factors for eviction, no rule has been defined for identifying a concerning level of eviction (or eviction filing) risk from the value of an associated factor. In this paper, we propose a novel approach to design such rules that mainly relies on (i) the shape of the probability distribution of the selected ACS features, (ii) whether the underlying factor is found to be a risk factor or a protective factor, (iii) the value of k (i.e., the desired percentage of hotspots in a county), and (iv) the characteristics of the county of each census tract. In the following paragraphs, we elaborate on the role of the aforementioned criteria in the design of our labeling functions.

We define a labeling function for each of our nine associated factors (as shown in Table 1) separately. Each of these labeling functions can abstain from providing labels for a data point if it is highly uncertain about the label of that data point. To this end, we analyze the probability distributions of our nine factors in each county separately and find out that all distributions are right-skewed (similar to

Figure 6.2), where the distribution's right tail is longer than its tail on the left side. Therefore, the data points that fall on the left-hand side of the probability distribution look somewhat similar to each other with respect to that selected factor. This piece of evidence has motivated us to design labeling functions that abstain from labeling the data points that fall on the left-hand side.

Next, we need to decide on the **polarity** of each labeling function, which refers to the type of labels that it can assign (e.g., in a binary classification problem, the polarity can be any of the following: Positive, Negative, or {Positive, Negative}). The polarity of each labeling function is defined as follows: A labeling function corresponding to a risk factor only assigns positive labels, and similarly, the one corresponding to a protective factor only assigns negative labels. We made this decision because when a factor is known to be a risk factor (resp. protective factor) for the prevalence of eviction, it is positively correlated with the higher (resp. lower) number of eviction. Thus, a larger (resp. smaller) value of this factor provides a signal on a larger (resp. smaller) number of evictions and eviction filings in that region. The polarity of our labeling functions is given in Table 6.1.

Furthermore, we need to specify the exact value of the threshold (shown in Figure 6.2) to complete the definition of our labeling functions. Since we want to find the top-$k\%$ hotspots of each county and both high precision and recall are equally important in this domain, the threshold for factor f and county c is defined as the $\lceil \frac{k \times n}{100} \rceil^{th}$ largest value of that factor among census tracts in county c, where n refers to the total number of census tracts in county c. As a result, each labeling function labels about $k\%$ of data points. Figure 6.2 summarizes our schema for defining labeling functions. Next, we explain how to integrate these noisy signals to assign (at most) one probabilistic label to each data point.



(a) Risk factor F_r (b) Protective factor F_p

Figure 6.2: The proposed approach for defining labeling functions.

Probabilistic Label Generation. Each labeling function provides a signal, with unknown accuracy, regarding the label of each data point. Now, we need to integrate those signals to generate (at most) one label per data point. One simple approach is to take the majority vote, however, due to some potential correlations between selected factors, majority voting might result in the "double counting" issue [126]. Therefore, we use Snorkel [108] to integrate the outputs of our labeling functions. To produce a probabilistic label, Snorkel learns a generative model (over labeling functions) that (1) models the correlations between labeling functions, and (2) estimates their accuracy (through examining the overlaps/conflicts in their output) during learning [108]. *Please note that this step is done in an unsupervised manner and Snorkel does not utilize any ground-truth labels for integrating the outputs of labeling functions.* In section 6.6.2, we evaluate the gain of employing Snorkel rather than the majority voting approach.

Although the proposed labeling approach is suitable (i.e., fast, easy to compute, and inexpensive) for creating a labeled training set from a large unlabeled dataset, it has two weaknesses that limit its usability for identifying eviction filing hotspots directly: (i) It does not necessarily label all data points (i.e., census tracts) because the underlying labeling functions refrain from labeling a data point if they are highly uncertain (i.e., the total coverage is associated with the value of k). (ii) The algorithm relies on the ACS data, which is released with a delay of about two years, and hence, it cannot be used for monitoring the most recent situation as the input of the model is not available for the past two years. Next, we propose a neural network model that can label all data points using satellite imagery, which is available at a high temporal resolution.

6.5.2 The Hotspot Prediction Model

In this section, first, we explain our rationale for choosing satellite imagery as input. Then, we describe the architecture of our proposed neural network model that aims at identifying top-k% eviction filing hotposts from satellite imagery.

Rationale for the Use of Satellite Imagery. We choose satellite imagery as the input of our model mainly because of three reasons: (i) Past literature [127, 128] has shown that urban poverty can be identified using satellite imagery (e.g., urban

trees provide useful signals for identifying income inequalities, and distinguishing poor neighborhoods from the rich ones [129, 130]), and given the strong association between eviction and poverty, we believe that satellite imagery would be a suitable source of data for identifying eviction filing hotspots at the census tract level as well. (ii) We hope that our neural network model can identify signs of gentrification, which has been shown to be associated with eviction, from satellite imagery [131–133]. (iii) Satellite imagery is available at high spatial ($\sim$10 meters) and temporal ($\sim$10 days) resolutions, which makes it an appropriate source of data for monitoring eviction filing hotspots in a timely manner.

The Neural Network Model. We now describe the architecture of our hotspot prediction model (the component on the left-hand side of Figure 6.1). The model takes the satellite images of a census tract and its county (as well as the mask matrices) as input and predicts whether that census tract is among the top-k% hotposts of that county or not. This neural network model extends the idea behind the ResNet model [134], while considering the challenges and characteristics of this problem domain. In the following paragraphs, we elaborate on these challenges and how they are addressed in this model.

The first challenge is that different census tracts have various shapes and sizes, and hence, input images can have various sizes. To address this challenge, we take the following steps: (1) we set the width (and height) of each satellite image to the third quartile of the width (and height) of all satellite images, (2) we build a mask matrix for each image to distinguish valid pixels from invalid ones, and (3) we incorporate *Partial Convolutional Layer* [135, 136] into our instance of ResNet (*Partial ResNet)* to make sure that the result of convolution in each layer only depends on the valid pixels. *Partial ResNet* mainly consists of three residual blocks whose parameters are the same as the first three residual blocks in ResNet-18. Please note that the kernel initializer for all partial convolutional layers is set to *he_normal* [137].

Furthermore, as the hotspots are defined with respect to each county, the neural network should consider the characteristics of that county in the prediction process. To this end, we employ the feature concatenation approach [138]; i.e., we apply a CNN model (i.e., CNN_{county}) on the satellite images of a county and concatenate extracted features (i.e., the output of FC_1) with the output of our *Partial ResNet*

model. CNN_{county} applies a partial convolutional layer with 4 filters on each input image of a county, concatenates their outputs, and then, employs three partial convolutional layers with 16, 32, and 64 filters (respectively). The kernel size for all partial convolutional layers in CNN_{county} is set to (7×7). Also, FC_1 and FC_2 (represented in Figure 6.1) are fully-connected layers with 64 and 128 neurons (respectively) and the $ReLU$ activation function.

Finally, since the generated probabilistic labels could be noisy, we consider two loss functions in our experiments: (1) the binary cross-entropy loss function (Equation 6.1), which is commonly used under a small noise rate, and (2) the *Active Passive Loss (APL)* (Equation 6.2), which has been shown to be highly effective under a large noise rate [139]. APL is defined as the sum of Normalized Cross Entropy (the left term in Equation 6.2) and Reverse Cross Entropy [140] (the right term in Equation 6.2). In the following equations, p refers to the output probability of the neural network classifier and q denotes the ground truth.

$$-\sum_{K=0}^{1}(q(k|x) \times \log p(k|x)) \tag{6.1}$$

$$\frac{-\sum_{K=0}^{1} q(k|x) \log p(k|x)}{-\sum_{j=0}^{1}\sum_{K=0}^{1} q(y=j|x) \log p(k|x)} - \sum_{K=0}^{1}(p(k|x) \times \log q(k|x)) \tag{6.2}$$

6.6 Experimental Evaluation

In this section, we first describe the set-up and data preparation process. Then, we provide an empirical evaluation of the performance of our label generation approach. Finally, we conduct a comparison between the accuracy of WARNER and various baselines and assess the contributions of its components to the overall performance.

6.6.1 Set-up

We implemented our codes in Python and used the following packages/libraries: keras (v. 2.7.0), tensorflow (v. 2.7.0), pandas (v. 1.1.5), numpy (v. 1.19.5), and scikit-learn (v. 1.0.2). In our experiments, we utilize Adam [98] with a learning rate of 2×10^{-4}, β_1 of 0.9, and β_2 of 0.999 as the optimizer for training the neural network models. Also, the maximum number of epochs is 100 and the early

stopping technique [99] is used to stop the training process once the loss value on the validation set does not degrade after ten epochs.

6.6.2 Evaluation of Generated Labels

In this section, we evaluate the accuracy of the generated labels (under various conditions) by comparing them to the ground-truth labels available for Dallas county, TX. Table 6.2 compares the performance of our label generation approach against the majority voting technique with different choices of k ($k \in \{5, 10, 15\}$) and training regions. We make the best performance bold and report the percentage of increase (in the predictive performance) achieved by employing WARNER (in the best case) compared to the majority voting approach in the last row (i.e., Gain). Please note that all performance metrics are computed on the subset of the test set (i.e., the testing portion of Dallas data) labeled by all models. In our experiments, the set of data points labeled by our model is a superset of the set of data points labeled by majority voting.

According to the results, on average, the majority voting approach (which does not involve learning an ML-based model) achieves an AUC of 0.711, which could be an indicator of the good quality of our labeling functions. Further, in general, employing Snorkel leads to a considerable improvement against majority voting. In fact, on average, our model outperforms the majority voting approach by 21.8% in terms of AUC, which shows the value of employing Snorkel (compared to taking the majority vote) for integrating outputs of our labeling functions.

Additionally, we do a cross-region test to investigate the generalizability of our label generation approach; i.e., we train our model on the unlabeled data of other counties in TX (i.e., all Texas counties except Dallas County), and then, evaluate its performance on the testing portion of the Dallas data. As a result, it achieves an AUC of 0.866 (on average), which is higher than the average AUC of the model trained on the training portion of Dallas data (i.e., 0.843). This shows that our label generation approach could be generalized to different counties within the USA.

Finally, in spite of the high accuracy with different choices of k, the coverage of this label generation approach can change with the value of k (as the coverage of the underlying labeling functions changes with k); e.g., in total, about 77%, 59%,

Table 6.2: A comparison between the performance of our label generation model and majority voting.

Model	Training Region	$k = 5$ F1	$k = 5$ AUC	$k = 10$ F1	$k = 10$ AUC	$k = 15$ F1	$k = 15$ AUC	Avg. ($k \in \{5, 10, 15\}$) F1	Avg. ($k \in \{5, 10, 15\}$) AUC
Majority Vote	—	0.307	0.690	0.538	0.809	0.400	0.634	0.415	0.711
WARNER	Dallas County, TX	0.307	0.851	**0.666**	0.914	0.500	0.766	0.491	0.843
WARNER	Other counties in TX	**0.666**	**0.886**	0.533	**0.943**	**0.533**	**0.771**	**0.577**	**0.866**
	Gain (%)	116.9%	28.4%	23.7%	16.5%	33.2%	21.6%	39.0%	21.8%

and 35% of data points are labeled by at least one labeling function when k is equal to 15, 10, and 5, respectively. However, a low coverage does not result in a serious issue in our problem domain because unlabeled data can be collected easily.

6.6.3 Evaluation of the Hotspot Prediction Model

We conduct three sets of experiments to assess the effectiveness of WARNER for the task of top-k% hotspot prediction. First, we conduct a comparison between the accuracy of WARNER and several strong deep learning-based baseline models. Then, we investigate the impact of WARNER's components on the value of different performance metrics. Finally, we evaluate the potential of WARNER trained for a specific k (e.g., $k = 10$) to be generalized (easily) to other values of k (e.g., $k \in \{5, 15\}$).

Comparison with Baseline Models. In this set of experiments, we consider the following three baseline models: (1) A Convolutional Neural Network (CNN) model with four convolutional layers, (2) Partial-CNN that incorporates partial convolutional layers [135] into the CNN model, instead of the standard convolutional layer, and (3) ResNet-18 [134] which is a residual network with 18 layers. While CNN and ResNet-18 take masked satellite images as input, Partial-CNN takes satellite images and mask matrices as separate inputs as it can distinguish valid and invalid pixels. Further, the binary cross-entropy loss function is utilized for training all neural models and evaluated on the testing portion of the Dallas data (in the next section, we compare the effectiveness of cross-entropy with that of APL in our problem domain).

Table 6.3 shows the performance of WARNER and the aforementioned baselines for $k \in \{5, 10, 15\}$. The first three rows show the performance of baseline models

Table 6.3: An evaluation of the performance of WARNER and baseline models.

Model	Training Region	$k=5$		$k=10$		$k=15$		Avg. ($k \in \{5,10,15\}$)	
		F1	AUC	F1	AUC	F1	AUC	F1	AUC
CNN	Dallas County, TX	0.000	0.631	0.162	0.544	0.136	0.560	0.099	0.578
ResNet-18	Dallas County, TX	0.000	0.632	0.166	0.588	0.138	0.634	0.101	0.618
Partial-CNN	Dallas County, TX	0.000	0.639	0.208	0.596	0.200	**0.639**	0.136	0.624
Partial-CNN	Other counties in TX	0.000	0.580	0.117	0.528	0.181	0.638	0.099	0.582
WARNER	Other counties in TX	**0.083**	**0.650**	**0.222**	**0.644**	**0.250**	0.607	**0.185**	**0.633**

being trained on the ground-truth data of Dallas County in a fully-supervised manner. In addition, the fourth and fifth rows represent the performance of Partial-CNN and WARNER (respectively) being trained on the labels that our label generation approach produced for the data of other counties in Texas. According to the results, WARNER outperforms the best-performing fully-supervised model (i.e., Partial-CNN) by 36.0% and 1.4% (on average) in terms of F1 and AUC, respectively. Therefore, *although WARNER has not seen any data from Dallas County during the training phase, it works better than the best-performing fully-supervised baseline model trained on the training portion of the Dallas data, which has ground-truth labels.*

Further, we observe higher improvements when training WARNER and the best-performing baseline model (i.e., Partial-CNN) on the same training dataset. In fact, the results of training both WARNER and Partial-CNN on the data of other counties in TX (with generated labels) show that WARNER outperforms Partial-CNN by 86.8% and 8.7% (on average) in terms of F1 and AUC, respectively. Further, the performance of Partial-CNN decreases significantly (by 27.2% and 6.7% in F1 and AUC, respectively) when being trained on the data of other counties and evaluated on the testing portion of Dallas data. This observation could show the impact of considering the satellite imagery of a county in the decision-making process because different characteristics of various counties could mislead the network.

Additionally, comparing the accuracy of fully-supervised baselines, we see that incorporating partial convolutional layers into CNN improves F1 and AUC by 37.3% and 7.9% (on average), respectively. Also, employing residual learning with a deeper network (i.e., ResNet-18) leads to 2.0% and 6.9% improvement (on average) in terms of F1 and AUC, respectively.

Finally, we note that almost all deep learning-based models have an AUC of

over 0.6 in the task of predicting top-5% hotspots (with a significantly imbalanced dataset), which could show the models' capability in distinguishing positive samples from negative ones. However, achieving a high F1 (which is calculated using the threshold of 0.5 on the predicted probability of belonging to the positive class) is an extremely difficult task in this case and we plan to address that in our future research.

Ablation Study. We now investigate the effect of various components on the overall accuracy of WARNER. Table 6.4 shows the outcome of the ablation study on WARNER while assuming $k = 10$. According to the results, replacing partial convolutional layers with the standard convolutional layers (i.e., `WARNER-w-Conv`) results in 20.2% and 9.0% decrease in F1 and AUC, respectively. This observation suggests that simply masking the invalid pixels could pose significant challenges to the learning process of our neural network.

Further, we compare the suitability of the feature concatenation approach with that of a recent condition approach called *Feature-wise Linear Modulation (FiLM)* method [141]. In fact, three common approaches have been usually used for incorporating multiple signals into a model [138]: input concatenation, feature concatenation, and conditioning layer. As the size of county images differs a lot from the size of census tract images, input concatenation is not an appropriate choice in our case. However, we tried a conditioning layer approach (i.e., `WARNER-w-FiLMLayer`) called *Feature-wise Linear Modulation (FiLM)* [141]. `WARNER-w-FiLMLayer` takes the following steps: (i) for the i^{th} census tract, it extracts two sets of features (i.e., $a^j_{i,m}$ and $b^j_{i,m}$) from the output of FC_1 through applying linear layers, and then (ii) use them to influence the m^{th} feature map of the j^{th} convolution layer ($f^j_{i,m}$) in the Partial ResNet via the feature-wise affine transformation given in Equation 6.3 [141]. According to the results, in our problem domain, employing FiLM leads to 22.0% and 14.5% decrease in F1 and AUC, respectively. Thus, the use of a concatenation layer seems to be a more appropriate choice.

$$FiLM(f^j_{i,m}|a^j_{i,m}, b^j_{i,m}) = a^j_{i,m} \times f^j_{i,m} + b^j_{i,m} \tag{6.3}$$

Additionally, comparing the performance of `WARNER` with that of `WARNER-w-APLLoss` shows that incorporating either of the two mentioned loss functions, i.e., cross-entropy and APL, results in a similar predictive performance. Finally, to evaluate

the value of our label generation approach in the WARNER framework, we replace it with a naive label generation approach and train our neural network model on this newly labeled data. This naive approach works as follows: A census tract is among top-k% hotspot if it shows up among the top-k% census tracts of its county in terms of the total population. The experimental results show that, the use of this naive label generation approach, i.e., `WARNER-w-NaiveLabel`, leads to 31.0% and 14.1% decrease in F1 and AUC, respectively. This observation shows the key role of our label generation approach in the WARNER's architecture for building a more accurate eviction filing hotspot prediction model.

Table 6.4: The results of ablation study when $k = 10$.

Model	F1	AUC	Drop in AUC (%)
WARNER	0.222	0.644	—
WARNER-w-Conv	0.177	0.586	-9.006%
WARNER-w-FiLMLayer	0.173	0.550	-14.596%
WARNER-w-APLLoss	0.226	0.643	-0.001%
WARNER-w-NaiveLabel	0.153	0.553	-14.130%

Generalizability of WARNER to Various Values of k. In our previous experiments, we trained a separate model for each value of k because the task changes with the value of k; i.e., `if` $k_1 \neq k_2 \rightarrow p(y_{k=k_1}|x) \neq p(y_{k=k_2}|x)$ (in this formula, y_k shows the binary label of interest for different values of k). However, training a separate model for each k of interest could be time-consuming. To tackle this challenge, we propose to use a transfer learning approach [142] to be able to easily transfer knowledge from a single pre-trained WARNER to the target task. The transfer learning algorithm works as follows: First, we train a single model for a specific value of k. Then, we freeze all weight matrices, except the parameters of the last two layers (as a result of freezing, the knowledge can be transferred to the target task). Finally, we fine-tune the parameters of the last two layers using the training data of the target task.

Table 6.5 represents the results of fine-tuning a WARNER model trained with $k = 10$ (i.e., source task) to the target tasks of top-5% hotspot prediction and top-15% hotspot prediction (please note that the results of training a separate WARNER for each target task are given in parenthesis). This table shows that

employing the aforementioned transfer learning approach leads to comparable results. Thus, we can easily transfer knowledge from a pre-trained WARNER model (trained with $k = 10$, for example) to the task of top-k′ hotspot prediction with various values of k′ (e.g., k′ $\in \{5, 15\}$ in our experiments).

Table 6.5: An evaluation of the generalizability of a pre-trained WARNER (with $k = 10$) to the task of top-k′ hotspot prediction ($k' \in \{5, 15\}$).

k'	F1 (original)	AUC (original)
5	0.000 (0.083)	0.676 (0.650)
15	0.181 (0.250)	0.590 (0.607)

6.7 Real-World Use Case

One possible use case for WARNER is to help NGOs, policymakers, and/or federal agencies improve their resource allocation plans to enhance housing stability. In fact, various rental assistance programs (like Emergency Rental Assistance Program (ERAP) [114]) are being implemented to assist high-need renters. While being available nationwide, a big variability has been observed in the utilization of those financial resources across the USA; e.g., some regions have sent back part of the funding to the government as it was too much money for those regions [84]. One potential reason for this issue could be the limited understanding of the distribution of eviction filings at the local level, which, in turn, results from some existing obstacles to eviction data acquisition. As a result, in the absence of ground-truth eviction filing data, WARNER could serve as an ML-based assistant for informing actions across the USA. In other words, WARNER's predictions can be visualized as a map of hotspots similar to Figure 6.3. Figure 6.3 represents WARNER's prediction of the top 10% hotspots (shown in red) across Texas over a period of three years (from 2017 to 2019). Such a heatmap could assist NGOs and policymakers in (i) monitoring eviction filing hotspots over a period of time, (ii) identifying high-need areas, particularly hidden hotspots (i.e., actual hotspots from where no/limited reports have been received due to the aforementioned obstacles to data acquisition), and hence, (iii) distributing resources and funding more efficiently.

Regarding the potential benefits of eviction data tools for mitigating the eviction

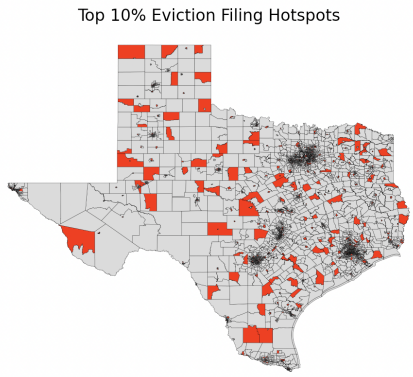


Figure 6.3: The WARNER's prediction regarding the top-10% eviction filing hotspots over a period of three years (from 2017 to 2019) across Texas. Hotspots and non-hotspots are shown with red and gray colors, respectively.

crisis, Ben Martin, an official at Texas Housers, mentioned that:

> "Eviction data tools are like the smoke that help us find the fire, and once we find the fire we can figure out what tools and resources to use to mitigate the problem."

He also described the importance of monitoring the eviction-related situation when eviction data is out of reach. Further, he provided more details on the contributions of such tools to improving the existing eviction diversion programs and related policies as follows.

> "Eviction data tools can contribute to these programs by, for instance, setting a baseline of need. Or, for a statewide ERA program, eviction data tools could help administrators identify areas of high need for targeted outreach."

The results of this discussion confirm (1) the significance of such ML-based tools for improving the existing eviction mitigation plans in the absence of eviction filing records (across a large region), and hence, (2) their high potential for making significant social impacts in the field.

6.8 Summary

This chapter proposed WARNER, which is a weakly-supervised ML- based framework for identifying eviction filing hotspots in US counties from satellite imagery in the absence of court records. In fact, first, it proposes a label generation approach that leverages sociological insights on the eviction crisis to label an unlabeled training dataset of satellite imagery. Then, relying on those generated labels, it built a neural network model for predicting eviction filing hospots from satellite imagery. To assess the performance of WARNER, it conducted various experiments using eviction filing data of Dallas County, TX. The experimental results show the suitability of the proposed label generation approach for this problem domain. Furthermore, WARNER outperforms multiple strong (fully-supervised) baseline models and its superior accuracy could be generalized to various counties within the US. In the absence of eviction filing records, the data-driven insights produced by WARNER could assist policymakers in distributing resources more efficiently and improving eviction mitigation programs.

Chapter 7
AI for Social Welfare of Housing-Insecure Low-Income Americans: Predicting Homeless Youth's Susceptibility to SUD

As a use case of ML, this chapter develops an ML model to identify homeless youth at-risk of Substance Use Disorder (SUD) with the goal of helping policymakers in their efforts at mitigating this urgent social problem [143].

7.1 Introduction

SUD refers to a pattern of harmful substance use (e.g., alcohol, marijuana, street and prescription opioids, stimulants, etc.) resulting in significant impairments [144]. Despite their negative side effects, sufferers continue to use these substances. SUD is a widespread and costly issue in the USA with abuse of tobacco, alcohol, and illicit drugs imposing over $740 billion each year [145]. In fact, about 19.7 million adults were reportedly suffering from SUD in 2017 [146]. More importantly, the SUD-related mortality rate has been increasing every year - it rose from 16 cases per 100,000 people (in 2002) to 27.5 cases per 100,000 (in 2015) [147].

In particular, SUD is more prevalent among the homeless youth population compared to the general public. For example, Busen and Engebretson [148] found that about 46% of their surveyed homeless youth suffered from SUD. Thus, any

attempt at tackling SUD at a national level crucially depends on our success at minimizing the rates of SUD among homeless youth.

Various programs and initiatives have been designed/implemented to tackle substance use/abuse among youth. One of them is the group-based intervention program [149,150]; in this intervention program, youth is split into multiple sub-groups, in which they get the opportunity to talk to peers, share their experiences, and hopefully, reinforce protective behaviors related to substance use. Such interactive programs are considered to be more effective to tackle substance use/abuse among young people than lecture-style programs because they take their peer's words more credible [150, 151]. However, the success of these programs is highly dependent on the sub-group formation strategy; e.g., assigning several high-risk individuals to the same sub-group could reinforce negative drug-using behaviors, a phenomenon which is known as *deviancy training* [149, 150]. Therefore, accurate information on the likelihood of each homeless youth suffering from SUD could potentially be helpful in effectively implementing such intervention programs among homeless youth.

As a step toward achieving this goal, we use a real-world dataset collected from ∼1,400 homeless youth from six states in the USA and build ML models to predict each homeless youth's susceptibility to SUD. Our best-performing model achieves an AUC of ∼0.85, which illustrates its high accuracy.

7.2 Related Work

In this section, we survey recent studies on alleviating the problems faced by the homeless population. These studies fall into two broad scientific areas: Artificial Intelligence and social science.

Artificial Intelligence Research. To the best of our knowledge, there had been no prior work on building and understanding models for predicting SUD among homeless youth. There has been a lot of interest in predicting substance use from social media data. Ding et al. [152] took advantage of several ML and text mining techniques to predict SUD. Hassanpour et al. [153] utilized a deep learning approach to predict the risk of substance use from Instagram profile data. However, the focus of these studies was mainly on the general population, and

thus, their results might not apply readily to homeless youth. Also, there is a growing body of work in AI on tackling problems faced by homeless youth. Yadav et al. [154, 155] and Rahmattalabi et al. [149, 156] focused on preventing Human Immunodeficiency Virus (HIV), substance abuse, and suicidal tendencies among the homeless youth population. However, most prior work in this space is concerned with finding prescriptive solutions, e.g., Yadav et al. [154] prescribe the selection of key influential homeless youth to spread awareness about HIV. On the other hand, this work aims to predict the susceptibility of homeless youth to SUD.

Social Science Research. Research with homeless populations is conducted in multiple social science disciplines with much of the work coming from sociology and psychology. While some of this work examines the effectiveness of interventions to address problems associated with homelessness, prior work also examines the experience of being homeless and how this relates to other aspects of an individual's life and well-being. Specifically, prior research investigates factors associated with an individual developing SUD. These factors can help identify at-risk individuals, which is important for outreach centers as they intervene in homeless populations. In particular, any form of child maltreatment (especially physical or sexual abuse) is shown to be a factor strongly associated with SUD [157–159]. While on the streets, trauma remains an associated factor for SUD irrespective of whether the individual witnessed a friend or loved one being victimized (indirect victimization), or if they had experienced the trauma themselves (direct victimization) [160]. Mental health disorders are also factors associated with SUD [157]. Other factors linked to SUD include demographic characteristics such as gender and age with young homeless men considered as one of the highest risk groups [158, 159]. Typically, prior studies in this space chose two or three groups of factors to investigate the level of their association with SUD. However, they did not focus on predicting the susceptibility of homeless youth to SUD, which is the focus of this chapter.

7.3 Dataset

In this work, we rely on a dataset collected from ∼1,400 homeless youth across six states in USA, namely California (CA), Arizona (AZ), Colorado (CO), Missouri (MO), Texas (TX), and New York (NY), from June 2016 until July 2017. Each

homeless youth was given a questionnaire to fill up, which consisted of questions about various topics including socio-demographic information (SD), criminal history (CH), sexual-risk behaviors (SR), victimization experiences (VE), gang involvement (GI), mental health characteristics (MH), and technology access (TA). Table 7.1 represents those topics along with the features corresponding to a couple of sample questions under these topics. This survey was approved by institutional review boards. For more information regarding the data collection procedures, please kindly refer to Barman-Adhikari et al. [161].

Topic	Feature	Explanation
SD	gender	Male, Female, Transgender, Gender queer, and other
CH	jail_homeless	Any jail or prison experiences since becoming unstably housed or homeless
	gunaccess	Having access to a gun or knowing how to access a gun easily
	avoid_police	Purposely avoiding situations that may expose you to interaction with police
SR	life_sexpartners	The number of sex partners in life
	last_sui_di	Drinking alcohol or using drugs before having sexual intercourse
	online_sexpart	Having sex with someone you met online
VE	ace	Experience of trauma and stress in childhood
	anyst_phy_vict	Any physical street victimization (e.g., assaulted with a weapon)
	witness_gun_di	Witnessing someone get attacked by a gun
GI	Juggalo_di	Ever been a Juggalo or a Juggalette?
MH	depression	The 9-item questionnaire (PHQ-9) is used to assess the level of depression
	ptsd	A 4-item questionnaire is used to measure PTSD
	perc_stress	Perceived stress during the past month
	unmet_ever	History of unmet mental health needs
	hospit_ever	History of staying in a hospital to treat mental health conditions
	medication_ever	Using medication to treat mental health conditions
	cope_8	How often do you use anger to get out of painful situations
	cope_9	How often do you use drugs or alcohol to deal with problems
TA	soc_media_prof	Having a profile on a social media site

Table 7.1: Summary of questionnaire topics with a couple of sample questions

Data Pre-processing. We pre-process the original dataset in two steps. First, as there are a lot of missing entries ($\sim$18.5%) in our dataset (as homeless youth could choose not to answer a question that made them feel uncomfortable), we used the MissForest algorithm [162], an off-the-shelf data imputation method to impute missing values in our dataset. Second, we apply feature standardization (i.e., Z-score normalization) to all features in our dataset. Finally, we randomly select 80% of samples as the training set and consider the remaining 20% as the test set. The class distribution in the training and test sets is set to be the same as

in the full dataset. At the end of this process, our data had 1,367 data points, each of which had 231 features and a binary label for predicting SUD.

7.4 SUD Prediction Model

We formulate the problem of predicting the susceptibility of homeless youth to SUD as a binary classification problem. To find the best performing model, we compared the accuracy of the following classification models: Logit, SVM [43], Classification And Regression Tree (CART) [163], Conditional Inference Forest (CForest) [164], XGBoost [42], AdaBoost [79], and an MLP with two hidden layers (the number of neurons in each hidden layer is half of that in the previous layer).

Table 7.2 compares the predictive performance of all our ML models across several widely used evaluation metrics. The rows in this table represent different classification algorithms and the columns represent different evaluation metrics (Accuracy, Precision, Recall, F1, and AUC). According to the results in Table 7.2, AdaBoost is the best-performing model in terms of all evaluation metrics. In particular, it achieves an AUC of 0.8546 which indicates its excellent class separation capability.

Model	Accuracy	Precision	Recall	F1	AUC
Logit	0.7032	0.5729	0.5789	0.5759	0.7776
SVM	0.7692	0.7285	0.5368	0.6181	0.8360
CART	0.7289	0.6779	0.4210	0.5194	0.6850
CForest	0.7728	0.7619	0.5052	0.6075	0.8507
XGBoost	0.7545	0.7000	0.5157	0.5939	0.8304
AdaBoost	**0.7985**	**0.7702**	**0.6000**	**0.6745**	**0.8546**
MLP	0.7362	0.6575	0.5052	0.5714	0.7010

Table 7.2: Performance of different ML models on predicting the susceptibility of homeless youth to SUD

In summary, experimental results show that it is indeed possible to train highly accurate ML models to predict the susceptibility of homeless youth to SUD. Next, we will conduct a feature importance analysis.

7.5 Feature Importance Analysis

In this section, first, we conduct an ablation study to investigate the relative importance of various sets of features on the overall performance of our prediction model. Then, we conduct an important analysis at the feature level.

Ablation Study. We now conduct a preliminary investigation into the relative importance of different sets of features in the predictive performance of our AdaBoost model. Specifically, we conduct an ablation study as follows: (1) we divide the features in our dataset into seven separate feature blocks (mentioned in section 7.3); each feature block consists of features related to a specific topic, e.g., one feature block ascertains involvement with gangs (GI), another block ascertains criminal history (CH), etc.; (2) we remove one feature block from the feature space (at a time), and then re-train an AdaBoost model on the remaining set of features; (3) finally, we report the percentage decrease in AUC values for our model.

Figure 7.1 shows the result of ablating different feature blocks. The X-axis shows the ablated feature block and the Y-axis shows the percentage decrease in AUC. According to the results, among all feature blocks, removing *mental health characteristics* (MH) leads to the greatest decrease in the model's predictive accuracy. At the same time, removing any of *sexual risk behavior* (SR), *gang involvement* (GI), and *victimization experiences* (VE) also leads to about 3% decrease in the model's AUC. In general, the results of our ablation study are consistent with a large body of literature that has established strong connections between mental health [165], sexual risk behavior [166], and victimization experiences [167] and SUD [168, 169].

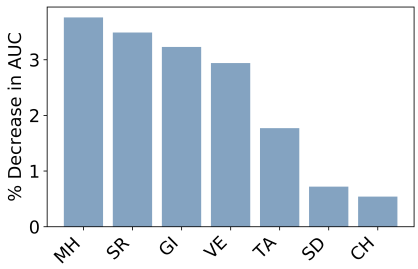


Figure 7.1: The results of ablation study

Importance of Individual Features. We now discover the subset of "important" features as follows: (1) we rank all features in our dataset based on their importance values; (2) starting from the most important feature, we add features one by one in the decreasing order of importance to the dataset and re-train a separate AdaBoost model (with only the restricted set of features). Figure 7.2 shows the AUC of the AdaBoost models trained with an increasing number of features. The X-axis shows the (increasing) number of features used to train the model and the Y-axis shows the AUC of the resulting AdaBoost model. This figure exhibits diminishing returns (in terms of the increases in AUC) beyond the addition of the 18 most important features in our dataset. Thus, we restrict our attention to these 18 features.

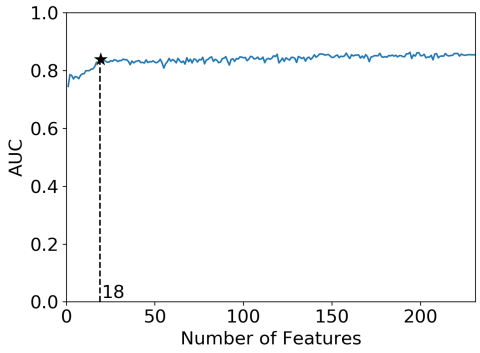


Figure 7.2: AUC of AdaBoost with different number of features

Figure 7.3 shows these 18 features ranked according to their importance values. The definition of these features can be seen in Table 7.1. Overall, we categorize these 18 features into three broad categories: environmental factors, psychological factors, and sexual-risk behaviors, and further analyze these categories in detail.

1. **Environmental Factors.** Our model finds some environmental factors important for predicting the susceptibility of homeless youth to SUD. In particular, perceived stress (`perc_stress`, Imp=0.662), adverse childhood experiences (`ace`, Imp=0.547), and some types of victimization experience (`anyst_phys_vict` and `witness_gun_di`, Average Imp=0.504) are among top-ranked features. This observation is consistent with existing literature as follows: (1) there is a lot of prior work which hypothesizes that homeless people's lifestyle (e.g., sleeping outside) increases the likelihood of experiencing victimization [170]. For example, Stewart et al. [171] show that ∼85% of

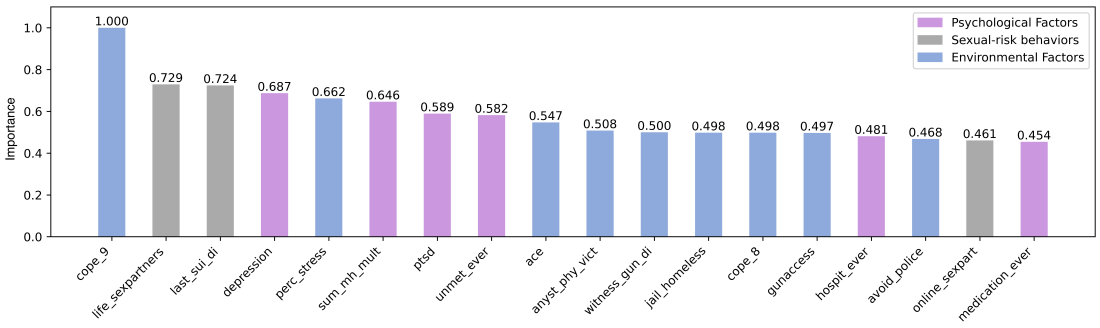


Figure 7.3: The importance value of 18 important features

the homeless population have experienced trauma and victimization. (2) These victimization experiences are shown to be significantly related to psychological distress and painful situations among youth [172]. (3) The importance of these factors along with factors related to coping strategies (`cope_8` and `cope_9`, average Imp=0.749) are consistent with prior work on SUD in homeless populations which shows that these youth self-medicate substances to alleviate the effect of painful situations and to cope with stressful situations [173].

Additionally, factors associated with law enforcement (`avoid_police` and `jail_homeless`, Average Imp=0.483) show up among important features. Intuitively, this makes sense because SUD involves the use of illicit substances (such as crack and cocaine), and an encounter with law enforcement could result in the individual being arrested and sent to jail. Even someone using a legal substance (e.g., alcohol) could be arrested for being intoxicated in public. Given the high punitive cost of engagement with law enforcement agencies, therefore, it is reasonable to expect that youth suffering from SUD would prefer to avoid encounters with law enforcement, or else they might end up in jail at some point during their time on the streets.

2. **Psychological Factors.** Our model finds some psychological factors important as well. In particular, certain mental health disorders (`ptsd` and `depression`, Average Imp=0.638) and mental health needs (e.g., `unmet_ever`, `hospit_ever`, `medication_ever`, Average Imp=0.505) show up among important features. The importance of these factors makes sense because prior work [174] suggests that people struggling with PTSD self-medicate and use

substances to cope with PTSD symptoms. Furthermore, the simultaneous presence of PTSD and depression among important factors along with the victimization experiences feature block is consistent with prior work [175], which shows that the comorbidity of depression and PTSD are highly likely among adolescents with victimization experiences.

3. **Sexual-Risk Behaviors.** Our model finds some factors pertaining to sexual-risk behavior important for predicting the susceptibility of homeless youth to SUD. In particular, we observe that factors related to sex partners (`life_sexpartners` and `online_sexpart`, Average Imp=0.595) and using substances before sex are among the important features (`last_sui_di`, Imp=0.724). In general, this could be justified (to some extent) by the existing literature [176] on the relationship between drug use and sexual risk behaviors. In particular, they explained that sex partners of drug users are highly likely to use drugs, and in this case, factors pertaining to sexual risk behaviors could be related to SUD.

7.6 Limitations

This work has a few limitations, many of which stem from the dataset that we use. The nature of the homeless population necessitates some decisions that limit the claims we can make with this research. The youth population surveyed for this study was not randomly selected which makes it more difficult to generalize our results to the entire population of homeless youth. Our data also relies on self-report measures, which have their own set of limitations. With self-report data, participants may not be completely honest when responding to the survey. Circumstances in this study make this more likely because the questions in the survey related to different conditions that have a stigma, making it possible that the individual would give a more socially acceptable answer instead of truth. As such, it is possible that conditions like SUD are under-reported in this dataset.

7.7 Summary

This chapter takes an ML-based approach to help policymakers and practitioners in their efforts to mitigate the prevalence of SUD among homeless youth. In fact, it develops an ML model with high accuracy to predict homeless youth's susceptibility to SUD. Then, it conducts feature importance analysis to obtain further insights.

Chapter 8
Future Work

This chapter describes a few future research pathways.

Multidisciplinary Approach to Label Generation. The lack of ground-truth labels is an important obstacle to the development of deep learning algorithms. In the future, I plan to build upon the existing research (such as [126]) and study how to generate high-quality labels of different types automatically. In particular, I am interested in translating findings of prior work in other disciplines into the target variable of interest in order to facilitate incorporating domain expertise into the process in an inexpensive and efficient manner.

Robust ML under Real-World Inaccuracies/Noises. Recent advances in the area of adversarial ML provide a great opportunity to develop ML models that are robust to adversarial perturbations, especially, many of them focused on classification tasks such as [177–179]. However, in certain real-world situations, accurate regression algorithms are needed, and the level of noise/inaccuracy might not necessarily be low. Accordingly, in the future, I plan to build upon existing research and design algorithms (with particular attention to regression tasks) that are robust to such real-world noises/inaccuracies.

Bibliography

[1] Jesslyn Brown (2018), "Monitoring Vegetation Drought Stress ," https://www.usgs.gov/special-topics/monitoring-vegetation-drought-stress/science/monitoring-vegetation-drought-stress-0, [Online; accessed 28-January-2023].

[2] India TV News Desk (2020), "Council votes to protect property of evicted tenants," https://www.indiatvnews.com/news/india/locust-attack-madhya-pradesh-districts-619089, [Online; accessed 01-February-2023].

[3] ROATEN, M. (2018), "Council votes to protect property of evicted tenants," https://www.streetsensemedia.org/article/dc-council-anita-bonds-charles-allen-elissa-silverman-david-grosso-us-marsh#.Y9sZ6y1h1ao, [Online; accessed 01-February-2023].

[4] Dake, L. (2022), "Federal data confirms Oregon spike in homelessness," https://www.opb.org/article/2022/12/24/federal-data-confirms-spike-oregon-homelessness/, [Online; accessed 01-February-2023].

[5] Tabar, M., W. Jung, A. Yadav, B. Martin, and D. Lee (2023) "Forecasting the Number of Tenants At-Risk of Formal Eviction at High Spatial Resolution with Low Resolution Ground-Truth Labels," [working paper].

[6] Tabar, M., W. Jung, A. Yadav, O. Wilson Chavez, A. Flores, and D. Lee (2022) "WARNER: Weakly-Supervised Neural Network to Identify Eviction Filing Hotspots in the Absence of Court Records," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 3514–3523.

[7] Tadele, Z. (2017) "Raising crop productivity in Africa through intensification," *Agronomy*, **7**(1), p. 22.

[8] Gollin, D. (2014) *Smallholder agriculture in Africa: An overview and implications for policy*, International Institute for Environment and Development (IIED).

[9] Salami, A., A. B. Kamara, and Z. Brixiova (2010) *Smallholder Agriculture in East Africa: Trends, Constraints and Opportunities*, Working Papers Series N° 105 African Development Bank, Tunis, Tunisia.

[10] Goedde, L., A. Ooko-Ombaka, and G. Pais (2019), "Winning in Africa's agricultural market," `https://www.mckinsey.com/industries/agriculture/our-insights/winning-in-africas-agricultural-market`, [Online; accessed 3-January-2021].

[11] Tabar, M., D. Lee, D. P. Hughes, and A. Yadav (2022) "Mitigating Low Agricultural Productivity of Smallholder Farms in Africa: Time-Series Forecasting for Environmental Stressors," in *The 34th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-22)*.

[12] Beyene, A. (2014) *Small farms under stress play a huge role for Africa: smallholder agriculture and emerging global challenges*, Nordiska Afrikainstitutet.

[13] Jia, X., M. Wang, A. Khandelwal, A. Karpatne, and V. Kumar (2019) "Recurrent generative networks for multi-resolution satellite data: an application in cropland monitoring," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2628–2634.

[14] Chien, J.-T. and K.-T. Kuo (2017) "Variational Recurrent Neural Networks for Speech Separation," in *INTERSPEECH*, pp. 1193–1197.

[15] He, J., D. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick (2019) "Lagging Inference Networks and Posterior Collapse in Variational Autoencoders," in *International Conference on Learning Representations (ICLR)*.

[16] Bowman, S. R., L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio (2016) "Generating Sentences from a Continuous Space," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 10–21.

[17] Lai, G., B. Li, G. Zheng, and Y. Yang (2018) "Stochastic wavenet: A generative latent variable model for sequential data," *arXiv preprint arXiv:1806.06116*.

[18] Pallikara Bahuleyan, H. (2018) *Natural Language Generation with Neural Variational Models*, Master's thesis, University of Waterloo.

[19] Yang, Z., Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick (2017) "Improved Variational Autoencoders for Text Modeling Using Dilated Convolutions," in *International conference on machine learning*, pp. 3881–3890.

[20] Desmond, M. and R. Kimbro (2015) "Eviction's Fallout: Housing, Hardship, and Health," *Social Forces*, **94**.

[21] Eggers, F. J., F. Moumen, and I. Econometrica (2010) "Investigating Very High Rent Burdens Among Renters in the American Housing Survey," *U.S. Department of Housing and Urban Development, Washington, DC*, **94**.

[22] Himmelstein, G. and M. Desmond (2021) "Association of eviction with adverse birth outcomes among women in Georgia, 2000 to 2016," *JAMA pediatrics*, **175**(5), pp. 494–500.

[23] Fan, J., D. Chen, J. Wen, Y. Sun, and C. P. Gomes (2022) "Monitoring Vegetation From Space at Extremely Fine Resolutions via Coarsely-Supervised Smooth U-Net," *arXiv preprint arXiv:2207.08022*.

[24] Chung, J., K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio (2015) "A recurrent latent variable model for sequential data," *Advances in neural information processing systems*, **28**, pp. 2980–2988.

[25] Harvey, C., Z. L. Rakotobe, N. S. Rao, R. Dave, H. Razafimahatratra, R. Rabarijohn, H. Rajaofara, and J. L. MacKinnon (2014) "Extreme vulnerability of smallholder farmers to agricultural risks and climate change in Madagascar," *Philosophical Transactions of the Royal Society B: Biological Sciences*, **369**(1639), p. 20130089.

[26] Shimeles, A., A. Verdier-Chouchane, and A. Boly (2018) "Introduction: understanding the challenges of the agricultural sector in Sub-Saharan Africa," in *Building a Resilient and Sustainable Agriculture in Sub-Saharan Africa*, Springer, pp. 1–12.

[27] Del Grosso, S., W. Parton, T. Stohlgren, D. Zheng, D. Bachelet, S. Prince, K. Hibbard, and R. Olson (2008) "Global potential net primary production predicted from vegetation class, precipitation, and temperature," *Ecology*, **89**(8), pp. 2117–2126.

[28] Sun, J. and W. Du (2017) "Effects of precipitation and temperature on net primary productivity and precipitation use efficiency across China's grasslands," *GIScience & Remote Sensing*, **54**(6), pp. 881–897.

[29] Zhang, S., R. Zhang, T. Liu, X. Song, and M. Adams (2017) "Empirical and model-based estimates of spatial and temporal variations in net primary productivity in semi-arid grasslands of Northern China," *PLoS ONE*, **12**(11), p. e0187678.

[30] Feng, K. and J. Tian (2021) "Forecasting reference evapotranspiration using data mining and limited climatic data," *European Journal of Remote Sensing*, **54**(sup2), pp. 363–371.

[31] Landeras, G., A. Ortiz-Barredo, and J. J. López (2009) "Forecasting weekly evapotranspiration with ARIMA and artificial neural network models," *Journal of irrigation and drainage engineering*, **135**(3), pp. 323–334.

[32] Alves, W., G. Rolim, and L. E. Aparecido (2017) "Reference Evapotranspiration Forecasting by Artificial Neural Network Models," *Engenharia Agrícola*, **37**, pp. 1116–1125.

[33] Izadifar, Z. (2010) *Modeling and Analysis of Actual Evapotranspiration using Data Driven and Wavelet Techniques*, Master's thesis, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

[34] Hochreiter, S. and J. Schmidhuber (1997) "Long short-term memory," *Neural computation*, **9**(8), pp. 1735–1780.

[35] Hyndman, R. J. and Y. Khandakar (2008) "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, **27**(3), pp. 1–22.

[36] Livera, A. M. D., R. J. Hyndman, and R. D. Snyder (2011) "Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing," *Journal of the American Statistical Association*, **106**(496), pp. 1513–1527.

[37] Zhang, L., C. Aggarwal, and G.-J. Qi (2017) "Stock price prediction via discovering multi-frequency trading patterns," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2141–2149.

[38] Food and Agriculture Organization of the United Nations (2018) *WaPOR Database Methodology: Level 2. Remote Sensing for Water Productivity Technical Report: Methodology Series*, FAO, Rome.

[39] Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998) *Crop evapotranspiration : guidelines for computing crop water requirements*, FAO.

[40] Roelofsen, P. (2018) *Time series clustering*, Master's thesis, Vrije Universiteit Amsterdam, Netherlands.

[41] Breiman, L. (2001) "Random Forests," *Machine learning*, **45**(1), pp. 5–32.

[42] Chen, T. and C. Guestrin (2016) "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

[43] Cortes, C. and V. Vapnik (1995) "Support-Vector Networks," *Machine Learning*, **20**(3), pp. 273–297.

[44] Bai, S., J. Z. Kolter, and V. Koltun (2018) "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*.

[45] Kingma, D. P. and M. Welling (2014) "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations (ICLR)*.

[46] Rezende, D. J., S. Mohamed, and D. Wierstra (2014) "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," in *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML)*, pp. 1278–1286.

[47] Dacrema, M. F., P. Cremonesi, and D. Jannach (2019) "Are we really making much progress? A worrying analysis of recent neural recommendation approaches," in *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109.

[48] Kisekka, I., K. W. Migliaccio, M. D. Dukes, B. Schaffer, J. H. Crane, H. K. Bayabil, and S. M. Guzman (2019), "Evapotranspiration-Based Irrigation Scheduling for Agriculture," `https://edis.ifas.ufl.edu/pdffiles/AE/AE45700.pdf`, [Online; accessed 12-January-2021].

[49] Brouwer, C., K. Prins, and M. Heibloem (1989) *Irrigation Water Management: Irrigation Scheduling*, Food and Agriculture Organization of the United Nations.

[50] Tabar, M., J. Gluck, A. Goyal, F. Jiang, D. Morr, A. Kehs, D. Lee, D. P. Hughes, and A. Yadav (2021) "A PLAN for Tackling the Locust Crisis in East Africa: Harnessing Spatiotemporal Deep Models for Locust Movement Forecasting," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3595–3604.

[51] Food and Agriculture Organization of the United Nations (2020), "Massive, border-spanning campaign needed to combat locust upsurge in East Africa," `http://www.fao.org/news/story/en/item/1257973/icode/`, [Online; accessed 2-February-2021].

[52] The World Bank (2020), "THE WORLD BANK GROUP AND THE LOCUST CRISIS," https://www.worldbank.org/en/topic/the-world-bank-group-and-the-desert-locust-outbreak, [Online; accessed 2-February-2021].

[53] Food and Agriculture Organization of the United Nations (2015), "FAO Desert Locust Information Service (DLIS)," http://www.fao.org/3/a-i4353e.pdf, [Online; accessed 2-February-2021].

[54] LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989) "Backpropagation Applied to Handwritten Zip Code Recognition," **1**(4), pp. 541–551.

[55] Food and Agriculture Organization of the United Nations and World Meteorological Organization (1965) *Meteorology and the Desert Locust: Proceedings of the WMO/FAO Seminar on Meteorology and the Desert Locust, Tehran, 25 Nov. - 11 Dec., 1963*, Publication, Secretariat of WMO.

[56] World Meteorological Organization and Food and Agriculture Organization of the United Nations (2016) *Weather and Desert Locusts*, *Tech. Rep. WMO-No. 1175*, World Meteorological Organization (WMO).

[57] Gómez, D., P. Salvador, J. Sanz, and J. L. Casanova (2020) "Modelling desert locust presences using 32-year soil moisture data on a large-scale," *Ecological Indicators*, **117**, p. 106655.

[58] Gómez, D., P. Salvador, J. Sanz, C. Casanova, D. Taratiel, and J. Casanova (2018) "Machine learning approach to locate desert locust breeding areas based on ESA CCI soil moisture," *Journal of Applied Remote Sensing*, **12**.

[59] Ye, S., S. Lu, X. Bai, and J. Gu (2020) "ResNet-Locust-BN Network-Based Automatic Identification of East Asian Migratory Locust Species and Instars from RGB Images," *Insects*, **11**(8).

[60] Kimathi, E., H. Tonnang, S. Subramanian, K. Cressman, E. Abdel-Rahman, M. Tesfayohannes, S. Niassy, B. Torto, T. Dubois, C. Tanga, M. Berresaw, S. Ekesi, D. Mwangi, and S. Kelemu (2020) "Prediction of breeding regions for the desert locust Schistocerca gregaria in East Africa," *Scientific Reports*, **10**.

[61] Space in Africa (2021), "AI Tool Will Help African Farmers Fight Locusts Using A Warning And Prediction System," https://africanews.space/ai-tool-will-help-african-farmers-fight-locusts-using-a-warning-and-prediction-system, [Online; accessed 2-February-2021].

[62] Yao, H., F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li (2018) "Deep multi-view spatial-temporal network for taxi demand prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 2588–2595.

[63] Yao, H., X. Tang, H. Wei, G. Zheng, and Z. Li (2019) "Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), pp. 5668–5675.

[64] Kumar, S., C. Peters-Lidard, Y. Tian, P. Houser, J. Geiger, S. Olden, L. Lighty, J. Eastman, B. Doty, P. Dirmeyer, J. Adams, K. Mitchell, E. Wood, and J. Sheffield (2006) "Land information system: An interoperable framework for high resolution land surface modeling," *Environmental Modelling & Software*, **21**(10), pp. 1402–1415.

[65] McNally, A., K. Arsenault, S. Kumar, S. Shukla, P. Peterson, S. Wang, C. Funk, C. Peters-Lidard, and J. Verdin (2017) "A land data assimilation system for sub-Saharan Africa food and water security applications," *Scientific Data*, **4**.

[66] Case, J., J. Mungai, V. Sakwa, B. Zavodsky, J. Srikishen, A. Limaye, and C. Blankenship (2016) "Transitioning Enhanced Land Surface Initialization and Model Verification Capabilities to the Kenya Meteorological Department (KMD)," *American Meteorological Society Fall Meeting*.

[67] ISRIC - World Soil Information (2020), "SoilGrids250m 2.0 - Sand content," `https://data.isric.org/geonetwork/srv/eng/catalog.search#/metadata/713396fa-1687-11ea-a7c0-a0481ca9e724`, [Online; accessed 2-February-2021].

[68] Funk, C. C., P. J. Peterson, M. F. Landsfeld, D. H. Pedreros, J. P. Verdin, J. D. Rowland, B. E. Romero, G. J. Husak, J. C. Michaelsen, A. P. Verdin, et al. (2014) "A quasi-global precipitation time series for drought monitoring," *US Geological Survey data series*, **832**(4), pp. 1–12.

[69] U.S. Geological Survey (2013), "GMTED2010 Viewer," `https://topotools.cr.usgs.gov/gmted_viewer/viewer.htm`, [Online; accessed 2-February-2021].

[70] NASA (2021), "POWER Data Access Viewer," https://power.larc.nasa.gov/data-access-viewer/, [Online; accessed 2-February-2021].

[71] Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, B. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph (1996) "The NCEP/NCAR 40-Year Reanalysis Project." *Bulletin of the American Meteorological Society*, **77**(3), pp. 437–472.

[72] Food and Agriculture Organization of the United Nations (2021), "WaPOR: The FAO portal to monitor WAter Productivity through Open access of Remotely sensed derived data," https://wapor.apps.fao.org/catalog/1, [Online; accessed 10-January-2021].

[73] Le Gall, M., R. Overson, and A. Cease (2019) "A Global Review on Locusts (Orthoptera: Acrididae) and Their Interactions With Livestock Grazing Practices," *Frontiers in Ecology and Evolution*, **7**, p. 263.

[74] Kaastra, I. and M. Boyd (1996) "Designing a neural network for forecasting financial and economic time series," *Neurocomputing*, **10**(3), pp. 215–236.

[75] Cao, L.-J. and F. E. H. Tay (2003) "Support vector machine with adaptive parameters in financial time series forecasting," *IEEE Transactions on neural networks*, **14**(6), pp. 1506–1518.

[76] Falessi, D., J. Huang, L. Narayana, J. F. Thai, and B. Turhan (2020) "On the need of preserving order of data when validating within-project defect classifiers," *Empirical Software Engineering*, **25**(6), pp. 4805–4830.

[77] Forman, G. and M. Scholz (2010) "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," *Acm Sigkdd Explorations Newsletter*, **12**(1), pp. 49–57.

[78] Kingma, D. P. and J. Ba (2015) "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations (ICLR)*.

[79] Freund, Y. and R. E. Schapire (1997) "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, **55**(1), pp. 119–139.

[80] Tabar, M., W. Jung, A. Yadav, O. W. Chavez, A. Flores, and D. Lee (2022) "Forecasting the Number of Tenants At-Risk of Formal Eviction: A Machine Learning Approach to Inform Public Policy," in *Proceedings of the*

31st International Joint Conference on Artificial Intelligence, IJCAI 2022, pp. 5178–5184.

[81] Desmond, M. (2012) "Eviction and the reproduction of urban poverty," *American journal of sociology*, **118**(1), pp. 88–133.

[82] The Eviction Lab (2018), "The Eviction Lab," [Online; accessed November 2023].
URL https://evictionlab.org/

[83] Gromis, A. (2019) "Eviction: Intersection of poverty, inequality, and housing," *Princeton, NJ: Princeton University, Eviction Lab*.

[84] PBS NewsHour (2022), "States clash over rental assistance as the federal government reallocates funds," Accessed: August 2022.
URL https://www.pbs.org/newshour/nation/states-clash-over-rental-assistance-as-the-federal-government-reallocates-f

[85] Bieretz, B., K. Burrowes, and E. Bramhall (2020), "Getting Landlords and Tenants to Talk: The Use of Mediation in Eviction," [Online; accessed November 2021].
URL https://www.urban.org/sites/default/files/publication/101991/getting-landlords-and-tenants-to-talk_3.pdf

[86] van Laere, I. R., M. A. de Wit, and N. S. Klazinga (2009) "Pathways into homelessness: recently homeless adults problems and service use before and after becoming homeless in Amsterdam," *BMC public health*, **9**(1), pp. 1–9.

[87] Desmond, M., W. An, R. Winkler, and T. Ferriss (2013) "Evicting children," *Social Forces*, **92**(1), pp. 303–327.

[88] Desmond, M. and C. Gershenson (2017) "Who gets evicted? Assessing individual, neighborhood, and network factors," *Social science research*, **62**, pp. 362–377.

[89] Montgomery, A. E., M. Cusack, D. Szymkowiak, J. Fargo, and T. O'toole (2017) "Factors contributing to eviction from permanent supportive housing: Lessons from HUD-VASH," *Evaluation and Program Planning*, **61**, pp. 55–63.

[90] Stenberg, S.-Å., L. Brännström, C. Lindberg, and Y. B. Almquist (2020) "Risk Factors for Housing Evictions: Evidence from Panel Data," *European Journal of Homelessness _ Volume*, **14**(2_).

[91] Hatch, M. E. and J. Yun (2021) "Losing your home is bad for your health: Short-and medium-term health effects of eviction on young adults," *Housing Policy Debate*, **31**(3-5), pp. 469–489.

[92] Greiner, D. J., C. W. Pattanayak, and J. Hennessy (2012) "The limits of unbundled legal assistance: a randomized study in a Massachusetts district court and prospects for the future," *Harv. L. rev.*, **126**, p. 901.

[93] Ye, T., R. Johnson, S. Fu, J. Copeny, B. Donnelly, A. Freeman, M. Lima, J. Walsh, and R. Ghani (2019) "Using machine learning to help vulnerable tenants in new york city," in *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 248–258.

[94] Tan, J. (2020) *Using machine learning to identify populations at high risk for eviction as an indicator of homelessness*, Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

[95] Puckett, A., L. M. Renner, and K. S. Slack (2002) *Trends in homelessness & housing insecurity*, *Tech. rep.*, Northwestern University.

[96] Bassuk, E. L., J. C. Buckner, L. F. Weinreb, A. Browne, S. S. Bassuk, R. Dawson, and J. N. Perloff (1997) "Homelessness in female-headed families: childhood and adult risk and protective factors." *American journal of public health*, **87**(2), pp. 241–248.

[97] Desmond, M., A. Gromis, L. Edmonds, J. Hendrickson, K. Krywokulski, L. Leung, and A. Porton (2018) *Eviction Lab Methodology Report: Version 1.0*, *Tech. rep.*, Princeton: Princeton University, Princeton, NJ, `www.evictionlab.org/methods`.

[98] Kingma, D. P. and J. Ba (2014) "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*.

[99] Prechelt, L. (1996) "Early Stopping-But When?" in *Neural Networks: Tricks of the Trade*.

[100] Hilt, D. E. and D. W. Seegrist (1977) *Ridge, a computer program for calculating ridge regression estimates*, vol. 236, Department of Agriculture, Forest Service, Northeastern Forest Experiment Station.

[101] Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017) "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, **30**.

[102] Arik, S. Ö. and T. Pfister (2021) "TabNet: Attentive Interpretable Tabular Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 6679–6687.

[103] Cho, K., B. Van Merriënboer, D. Bahdanau, and Y. Bengio (2014) "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*.

[104] Tan, J. (2020) *Using machine learning to identify populations at high risk for eviction as an indicator of homelessness*, Master's thesis, MIT, Cambridge, MA.

[105] Xie, M., N. Jean, M. Burke, D. Lobell, and S. Ermon (2016) "Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, p. 3929–3935.

[106] Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016) "Combining satellite imagery and machine learning to predict poverty," *Science*, **353**(6301), pp. 790–794.

[107] Carbonneau, M.-A., V. Cheplygina, E. Granger, and G. Gagnon (2018) "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, **77**, pp. 329–353.

[108] Ratner, A. J., S. H. Bach, H. R. Ehrenberg, and C. Ré (2017) "Snorkel: Fast training set generation for information extraction," in *Proceedings of the 2017 ACM international conference on management of data*, pp. 1683–1686.

[109] Desmond, M. and T. Shollenberger (2015) "Forced displacement from rental housing: Prevalence and neighborhood consequences," *Demography*, **52**(5), pp. 1751–1772.

[110] Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender (2005) "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96.

[111] Burges, C. J. (2010) *From RankNet to LambdaRank to LambdaMART: An Overview*, *Tech. Rep. MSR-TR-2010-82*.
URL `https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/`

[112] Benfer, E. A., S. J. Greene, and M. Hagan (2020) "Approaches to eviction prevention," *Available at SSRN 3662736*.

[113] Treskon, M., S. Greene, O. Fiol, and A. Junod (2021) "Eviction prevention and diversion programs," *Washington, DC: Urban Institute*.

[114] U.S. Department of the Treasury (2021), "Emergency Rental Assistance Program," https://home.treasury.gov/policy-issues/coronavirus/assistance-for-state-local-and-tribal-governments/emergency-rental-assistance-program, [Online; accessed 4-February-2022].

[115] Office of Policy Development & Research (2021) *Report to Congress on the Feasibility of Creating a National Evictions Database*, *Tech. rep.*, U.S. Department of Housing and Urban Development.

[116] Desmond, M. (2015) "Unaffordable America: Poverty, housing, and eviction," *Fast Focus: Institute for Research on Poverty*, **22**(22), pp. 1–6.

[117] Rutan, D. Q. and M. Desmond (2021) "The concentrated geography of eviction," *The ANNALS of the American Academy of Political and Social Science*, **693**(1), pp. 64–81.

[118] LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989) "Backpropagation Applied to Handwritten Zip Code Recognition," **1**(4), pp. 541–551.

[119] Head, A., M. Manguin, N. Tran, and J. E. Blumenstock (2017) "Can Human Development be Measured with Satellite Imagery?" in *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, pp. 1–11.

[120] Ayush, K., B. Uzkent, K. Tanmay, M. Burke, D. Lobell, and S. Ermon (2021) "Efficient Poverty Mapping from High Resolution Remote Sensing Images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 12–20.

[121] Ayush, K., B. Uzkent, M. Burke, D. Lobell, and S. Ermon (2020) "Generating Interpretable Poverty Maps using Object Detection in Satellite Images," in *International Joint Conferences on Artificial Intelligence*.

[122] Lee, J., D. Grosz, B. Uzkent, S. Zeng, M. Burke, D. Lobell, and S. Ermon (2021) "Predicting Livelihood Indicators from Community-Generated Street-Level Imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 268–276.

[123] Hu, W., J. H. Patel, Z.-A. Robert, P. Novosad, S. Asher, Z. Tang, M. Burke, D. Lobell, and S. Ermon (2019) "Mapping missing population in rural India: A deep learning approach with satellite imagery," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 353–359.

[124] Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020) "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature communications*, **11**(1), pp. 1–11.

[125] Tabar, M., W. Jung, A. Yadav, O. W. Chavez, A. Flores, and D. Lee (2022) "Forecasting the Number of Tenants At-Risk of Formal Eviction: A Machine Learning Approach to Inform Public Policy," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 5178–5184.

[126] Ratner, A., S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré (2020) "Snorkel: Rapid training data creation with weak supervision," *The VLDB Journal*, **29**(2), pp. 709–730.

[127] Lin, L., L. Di, C. Zhang, L. Guo, and Y. Di (2021) "Remote Sensing of Urban Poverty and Gentrification," *Remote Sensing*, **13**(20), p. 4022.

[128] Li, G., Z. Cai, Y. Qian, and F. Chen (2021) "Identifying Urban Poverty Using High-Resolution Satellite Imagery and Machine Learning Approaches: Implications for Housing Inequality," *Land*, **10**(6), p. 648.

[129] Zhu, P. and Y. Zhang (2008) "Demand for urban forests in United States cities," *Landscape and urban planning*, **84**(3-4), pp. 293–300.

[130] Per Square Mile (2012), "Urban trees reveal income inequality," [Online; accessed 27-November-2021].
URL `https://persquaremile.com/2012/05/17/urban-trees-reveal-income-inequality/`

[131] Chum, A. (2015) "The impact of gentrification on residential evictions," *Urban Geography*, **36**(7), pp. 1083–1098.

[132] Marcuse, P. (2013) "Abandonment, gentrification, and displacement: the linkages in New York City," in *Gentrification of the City*, Routledge, pp. 169–193.

[133] Newman, K. and E. K. Wyly (2006) "The right to stay put, revisited: Gentrification and resistance to displacement in New York City," *Urban studies*, **43**(1), pp. 23–57.

[134] He, K., X. Zhang, S. Ren, and J. Sun (2016) "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

[135] Liu, G., F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro (2018) "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100.

[136] Liu, G., K. J. Shih, T.-C. Wang, F. A. Reda, K. Sapra, Z. Yu, A. Tao, and B. Catanzaro (2018) "Partial Convolution based Padding," in *arXiv preprint arXiv:1811.11718*.

[137] He, K., X. Zhang, S. Ren, and J. Sun (2015) "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

[138] AlBahar, B. and J.-B. Huang (2019) "Guided image-to-image translation with bi-directional feature transformation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9016–9025.

[139] Ma, X., H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey (2020) "Normalized loss functions for deep learning with noisy labels," in *International Conference on Machine Learning*, PMLR, pp. 6543–6553.

[140] Wang, Y., X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey (2019) "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330.

[141] Perez, E., F. Strub, H. De Vries, V. Dumoulin, and A. Courville (2018) "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

[142] Zhuang, F., Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He (2020) "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, **109**(1), pp. 43–76.

[143] Tabar, M., H. Park, S. Winkler, D. Lee, A. Barman-Adhikari, and A. Yadav (2020) "Identifying Homeless Youth At-Risk of Substance Use Disorder: Data-Driven Insights for Policymakers," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3092–3100.

[144] American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*, American Psychiatric Association.

[145] National Institute on Drug Abuse (2020), "Costs of Substance Abuse," www.drugabuse.gov/drug-topics/trends-statistics/costs-substance-abuse, [Online; accessed 14-June-2020].

[146] Substance Abuse and Mental Health Services Administration (2018) *Key Substance Use and Mental Health Indicators in the United States: Results from the 2017 National Survey on Drug Use and Health*, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.

[147] Snider, J. T., M. E. Duncan, M. R. Gore, S. Seabury, A. R. Silverstein, M. G. Tebeka, and D. P. Goldman (2019) "Association Between State Medicaid Eligibility Thresholds and Deaths Due to Substance Use Disorders," *JAMA Network Open*, **2**(4).

[148] Busen, N. H. and J. C. Engebretson (2008) "Facilitating risk reduction among homeless and street-involved youth," *Journal of the American Academy of Nurse Practitioners*, **20**(11), pp. 567–575.

[149] Rahmattalabi, A., A. Barman-Adhikari, P. Vayanos, M. Tambe, E. Rice, and R. Baker (2019) "Social Network Based Substance Abuse Prevention via Network Modification (A Preliminary Study)," *arXiv preprint arXiv:1902.00171*.

[150] Valente, T. W., B. R. Hoffman, A. Ritt-Olson, K. Lichtman, and C. A. Johnson (2003) "Effects of a social-network method for group assignment strategies on peer-led tobacco prevention programs in schools," *American journal of public health*, **93**(11), pp. 1837–1843.

[151] Black, D. R., N. S. Tobler, and J. P. Sciacca (1998) "Peer Helping/Involvement: An Efficacious Way to Meet the Challenge of Reducing Alcohol, Tobacco, and Other Drug Use Among Youth?" *Journal of School Health*, **68**(3), pp. 87–93.

[152] Ding, T., W. K. Bickel, and S. Pan (2017) "Multi-view unsupervised user feature embedding for social media-based substance use prediction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2275–2284.

[153] Hassanpour, S., N. Tomita, T. Delise, B. Crosier, and L. A. Marsch (2019) "Identifying substance use risk based on deep neural networks and Instagram social media data," *Neuropsychopharmacology*, **44**(3), pp. 487–494.

[154] Yadav, A., H. Chan, A. Xin Jiang, H. Xu, E. Rice, and M. Tambe (2016) "Using Social Networks to Aid Homeless Shelters: Dynamic Influence

Maximization under Uncertainty," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 740–748.

[155] Yadav, A., B. Wilder, E. Rice, R. Petering, J. Craddock, A. Yoshioka-Maxwell, M. Hemler, L. Onasch-Vera, M. Tambe, and D. Woo (2017) "Influence Maximization in the Field: The Arduous Journey from Emerging to Deployed Application," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 150–158.

[156] Rahmattalabi, A., P. Vayanos, A. Fulginiti, E. Rice, B. Wilder, A. Yadav, and M. Tambe (2019) "Exploring algorithmic fairness in robust graph covering problems," in *Advances in Neural Information Processing Systems*, pp. 15776–15787.

[157] Dietz, T. L. (2007) "Predictors of reported current and lifetime substance abuse problems among a national sample of U.S. homeless," *Substance Use & Misuse*, **42**, pp. 1745–1766.

[158] Tyler, K. A. and L. A. Melander (2015) "Child Abuse, Street Victimization, and Substance Use Among Homeless Young Adults," *Youth & Society*, **47**(4), pp. 502–519.

[159] Tyler, K., L. Kort-Butler, and A. Swendener (2014) "The Effect of Victimization, Mental Health, and Protective Factors on Crime and Illicit Drug Use Among Homeless Young Adults," *Violence and victims*, **29**(2), pp. 348–362.

[160] Thompson, S. J., K. Bender, K. M. Ferguson, and Y. Kim (2015) "Factors associated with substance use disorders among traumatized homeless youth," *Journal of social Work Practice in the Addictions*, **15**, pp. 66–89.

[161] Barman-Adhikari, A., H.-T. Hsu, D. Brydon, R. Petering, D. Santa Maria, S. Narendorf, J. Shelton, K. Bender, and K. Ferguson (2019) "Prevalence and correlates of nonmedical use of prescription drugs (NMUPD) among Young adults experiencing homelessness in seven cities across the United States," *Drug and Alcohol Dependence*, **200**, pp. 153–160.

[162] Stekhoven, D. J. and P. Buehlmann (2012) "MissForest–non-parametric missing value imputation for mixed-type data," *Bioinformatics*, **28**, pp. 112–118.

[163] Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984) *Classification and Regression Trees*, CRC press.

[164] Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan (2006) "Survival ensembles," *Biostatistics*, **7**(3), pp. 355–373.

[165] Heath, L. M., L. Laporte, J. Paris, K. Hamdullahpur, and K. J. Gill (2018) "Substance misuse is associated with increased psychiatric severity among treatment-seeking individuals with borderline personality disorders," *Journal of Personality Disorders*, **32**(5), pp. 694–708.

[166] Villalobos-Gallegos, L., M. E. Medina-Mora, C. Benjit, S. Ruiz-Velasco, C. Margis-Rodriguez, and R. Marin-Navarrete (2019) "Multidimensional patterns of sexual risk behavior and psychiatric disorders in men with substance use disorders," *Archives of Sexual Behavior*, **48**(2), pp. 599–607.

[167] Davis, J. P., E. R. Dworkin, J. Helton, J. Prindle, S. Patel, T. M. Dumas, and S. Miller (2019) "Extending poly-victimization theory: Differential effects of adolescents' experiences of victimization on substance use disorder diagnosis upon treatment entry," *Child Abuse & Neglect*, **89**, pp. 165–177.

[168] Ross, S. and E. Peselow (2012) "Co-Occurring Psychotic and Addictive Disorders," *Clinical neuropharmacology*, **35**, pp. 235–43.

[169] Kelly, T. M. and D. C. Daley (2013) "Integrated Treatment of Substance Use and Psychiatric Disorders," *Social work in public health*, **28**, pp. 388–406.

[170] Lee, B. A. and C. J. Schreck (2005) "Danger on the Streets: Marginality and Victimization Among Homeless People," *American Behavioral Scientist*, **48**(8), pp. 1055–1081.

[171] Stewart, A. J., M. Steiman, A. M. Cauce, B. N. Cochran, L. B. Whitbeck, and D. R. Hoyt (2004) "Victimization and Posttraumatic Stress Disorder Among Homeless Adolescents," *Journal of the American Academy of Child & Adolescent Psychiatry*, **43**(3), pp. 325–331.

[172] Hartinger-Saunders, R. M., B. Rittner, W. Wieczorek, T. Nochajski, C. M. Rine, and J. Welte (2011) "Victimization, psychological distress and subsequent offending among youth," *Children and Youth Services Review*, **33**(11), pp. 2375–2385.

[173] Tyler, K. A. and K. Johnson (2006) "Pathways in and out of substance use among homeless-emerging adults," *Journal of Adolescent Research*, **21**, pp. 133–157.

[174] Leeies, M., J. Pagura, J. Sareen, and J. M. Bolton (2010) "The use of alcohol and drugs to self-medicate symptoms of posttraumatic stress disorder," *Depression and anxiety*, **27**(8), pp. 731–736.

[175] Kilpatrick, D. G., K. J. Ruggiero, R. Acierno, B. E. Saunders, H. S. Resnick, and C. L. Best (2003) "Violence and risk of PTSD, major depression, substance abuse/dependence, and comorbidity: results from the National Survey of Adolescents," *Journal of consulting and clinical psychology*, **71**(4), pp. 692–700.

[176] Ross, M. W. and M. L. Williams (2001) "Sexual behavior and illicit drug use," *Annual review of sex research*, **12**(1), pp. 290–310.

[177] Sheikholeslami, F., A. Lotfi, and J. Z. Kolter (2021) "Provably robust classification of adversarial examples with detection," in *International Conference on Learning Representations*.

[178] Goodfellow, I., J. Shlens, and C. Szegedy (2015) "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations*.
URL `http://arxiv.org/abs/1412.6572`

[179] Apostolidis, K. D. and G. A. Papakostas (2021) "A Survey on Adversarial Deep Learning Robustness in Medical Image Analysis," *Electronics*, **10**(17).