A Regionally Generalized Machine Learning Framework Towards Census-Enabled Multi-Factor Non-Communicable Disease Analyses

Kevin Geng
Independent Researcher
Dublin, United States of America
kevin@ncdcare.org

Sudith Thota
Independent Researcher
Dublin, United States of America
sudith@ncdcare.org

Anish Kataria
Princeton University
Princeton, United States of America
anish@ncdcare.org

***Abstract*—In recent years, non-communicable diseases, diseases intransmittable through human interaction, have cemented themselves as a leading cause of death globally. These diseases are exacerbated by various social, economic and geographical factors within communities. Despite their resounding consequences on community health, the compilation of medical surveys is both costly and time consuming. This study uses a machine learning approach to predict the population proportion with Asthma, High Blood Pressure, Cancer, Diabetes, and Poor Mental Health within census tract populations in three states. Through an evaluation of 8 different regression models on their predictive capabilities of disease prevalence, the Cubist model held the strongest results for every disease, achieving considerably low relative mean value errors of 3.849%, 6.077%, 7.656%, 7.866%, and 5.531% towards each of the non-communicable diseases, respectively. Out of the four categories split from 41 input features- Sociodemographic, Economic, Environment/Geographic, and Environmental Justice-Sociodemographic features consistently were the highest contributors. A lightweight model with comparable accuracy to the full feature set (≤0.1% difference) was found to require only the 20 most important features from every Cubist model. This study provides insight towards machine learning models' capabilities in enhancing the United States' understanding of domestic epidemiology determinants.**

Keywords—non-communicable diseases, census tracts, community health outcomes, Cubist, machine learning

I. Introduction

The rising "invisible epidemic" [1], non-communicable diseases (NCDs)-diseases induced by human environment intransmissible through personal contact- have cemented themselves as a leading cause of death, accounting for over 41 million, or 74% of, global deaths annually [2]. These diseases fall under four primary classifications: cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes mellitus [1]. Given their resounding consequences towards global health, early detection and intervention is crucial in preventing acute disease contraction from advancing to chronic stages. Currently, key perpetrators of interest include sociodemographic determinants, or those that influence health, behaviors, and well-being of populations or environmental justice, which is the study of demographic-based influences affecting the proximity of health hazards, including power plants, railways, and contaminated water bodies. Existing medical health surveys, though accurate, are time-consuming and whose results are often deemed obsolete in a matter of years due to the rapidly shifting landscape of NCDs. Additionally, as wealth and social disparities widen across the United States, data reporting bias may prevail, skewing disease prediction data toward populations with greater access to healthcare.

Machine learning, first proposed by Arthur Samuel in his 1959 paper [3], offers a solution to traditional, precomputed algorithms by learning spatio-temporal relationships between dataset features, enabling a computer algorithm to derive its own constants and coefficients that best approximate continuous and discrete values. In the healthcare field, machine learning has seen a breadth of usages, including Support Vector Machines for breast cancer prediction and diagnosis [4], Deep Learning for pneumonia imaging classification [5], and Extreme Gradient Boosting for binary classification of neonatal mortality [6]. These models are tailored to specific tasks, with no one-size-fits-all solutions currently existing. While most existing machine learning techniques are outperformed by medical professionals, the rapid development and refinement of new technologies holds a promising future in machine learning working closer alongside professionals in detecting, and ultimately, preventing, various healthcare complications.

There exists research in the realm of machine learning for disease prediction. A research paper by Luo et al. [7] utilized linear, regularized, and decision tree models in order to predict the prevalence of 6 NCDs across all 50 United States, achieving a median Pearson correlation of 0.88. However, they noted potential biases due to non-standardized data collection, which hindered model predictive capabilities. Similarly, Feng et al. [8] conducted a study on 6 NCDs across 196 census tracts in Austin, Texas, using a variety of national and local datasets, including data from the United States Smart Location Database (SLD), Social Vulnerability Index (SVI), 500 Cities Project, and City of Austin's 311 service request data. The researchers found that the SVI dataset, in conjunction with the 311 data, performed best on their 6 measured NCDs. While the models performed reasonably well, some of their employed datasets contained data for only select regions, not providing coverage for a majority of cities across the United States. Thus, both studies highlight the need for diverse and standardized data samples to accurately predict health outcomes.

This study aims to provide a machine learning approach towards predicting and understanding the prevalences of five NCDs within communities: asthma, high blood pressure, cancer, diabetes, and poor mental health. This region of study was chosen due to its existing socio-demographic interest between recent devastating wildfires, wealth disparities among urban cities, and diverse topographies; a generally representative subset that could serve as a precursor for nationwide NCD evaluation.

II. Methodology

The complete machine learning workflow (Fig. 1) consists of data retrieval, data preprocessing, model selection, model fine tuning, model evaluation, and result visualization and interpretation, in that respective order. The iterative process between model fine tuning and model evaluation results in continuous feedback for algorithmic performance, ensuring the final model is optimal towards predicting all five output diseases. The careful selection of models based on past literature and predictive potential ensures that results are worthwhile for this application.

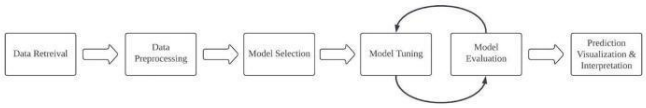

Fig. 1. Diagram of the employed machine learning workflow.

A. Data Retrieval and Feature Selection

The two datasets used were the United States Environmental Justice Index (EJI) and Smart Location Database (SLD). The EJI data, from the Center for Disease Control and Prevention, contains social, economic, demographic, and health outcome data from every census tract in the United States. The other dataset, the (SLD) from the US Environmental Protection Agency (EPA), contains geophysical, demographic, and economic data from census tracts in the United States. These census tracts, which vary both in area and population, allow for surveys and other population-based metrics to be compiled by the United States Census Bureau. The two databases were initially filtered down to the three target states by their FIPS codes, corresponding to 6, 41, and 53, respectively, for California, Oregon, and Washington. In the data, every census tract has a unique designation achieved through its state FIPS, county FIPS, and tract CE codes. Although the EJI dataset solely contained unique tracts, the SLD data contained numerous census tracts with multiple data points. For these tracts, weighted sums were required to be calculated for all target features. Given the varying population densities and areas of the 10,243 census tracts, only percentage-based features were selected as candidates for the filtered dataset. In total, 43 input and five output features were selected for the dataset (Table S1). Out of the input features, three features ("Pct_Wtr", "Pct_Land", "Pct_Unpr") were created through feature engineering existing geographical variables from the SLD, and relate to the proportion of land in a given census tract encompassed by water, land, and land legislatively unprotected from industrial development, respectively. With the complex nature of disease predictions, pulling multi-faceted data that describe a variety of the factors within a census tract is imperative. Thus, the input features were classed into four categories, with no external biases being attributed with any given category: Sociodemographic, Environment/Geographic, Environmental Justice, and Economic. The five predicted outputs include the risk of Asthma, High Blood Pressure, Cancer, Diabetes, and Poor Mental Health among a census tract's populations, all of which were retrieved from the EJI dataset.

B. Data Preprocessing

Both the input and output data were found to carry varying degrees of skew, which has been proven extensively to decrease machine learning model performance [9]. While models generally resistant to skew do exist, transformations were applied to all input features with an absolute skew coefficient greater than 0.5. Positive skews beyond the threshold had a normal Quantile Transformation applied, while negative skews beyond the threshold experienced a Yeo-Johnson transformation [10]. The two transformations' preservation of the rank order of features makes them extremely effective at preserving data integrity. While some features still carried varying degrees of skew following the transformations, the features generally became far more interpretable by the models. Following the input transformations, a pairwise correlation matrix was generated (Fig. S1), containing the degree of linear correlation between variables, with absolute values closer to 1 representing stronger correlations and value's signum representing their respective correlation (positive or negative). To reduce the risk of collinearity, which results in models that erroneously place excessive influence on highly correlated features, either of two variables was removed if their correlation coefficient exceeded 0.95. This resulted in the removal of two features ("D1B" and "D2A_WRKEMP"), reducing the dataset into 41 input features. The complete feature correlation matrix is included in the supplemental information.

The output variable distributions were also analyzed. Although outliers do exist, their distributions concur with the breadth of different socioeconomic and demographic compositions across the analyzed census tracts, and thus remain imperative to keep. While most of the outputs hovered around a positive skew coefficient of 0.5, the output prevalence of Cancer contained a moderately high skew coefficient of 0.92, potentially hindering model predictive abilities. Although prior machine learning studies have used the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOTER-GN) technique [11]. Its over-sampling results in higher-complexity models with larger datasets, increased training time, and loss of original data integrity. Thus, a y-transformation technique was used instead to construct a Gaussian output distribution. For this data, a Box-Cox transformation [12], proven to hold better performance and reliability than primitive (log, inverse, square-root) transformations [13], was used:

$$X = \begin{cases} \frac{X^{\lambda}-1}{\lambda}; & \lambda \neq 0 \\ \log(X); & \lambda = 0 \end{cases}$$

where lambda (λ) represents the power to exponentiate the value X by. The $\lambda = 0$ function can be derived as the function as the limit of the $\lambda \neq 0$ piecewise approaches $\lambda = 0$. For this

transformation, a larger absolute lambda equates to a greater enacted power transformation, creating a versatile transformation approach suitable for multiple varied skew levels. Additionally, the inclusion of the lambda value allows for the inverse Box Cox transformation to be applied to a transformed dataset, effectively returning it to its original state, a technique used in the subsequent sections to equate the evaluation of Cancer prevalence with other outputs. By applying a Box-Cox transformation, the right-skewed Cancer prevalence fell from a coefficient of 0.92 down to 0.04 (Fig. 2). To holistically and accurately represent our model, the census tract data was shuffled, and then broken into a train-validation and test split of 90/10, for 9218 train-validation and 1025 test rows; the extent of training data ensures the machine learning models are exposed to the largest majority of possible census tract cases.

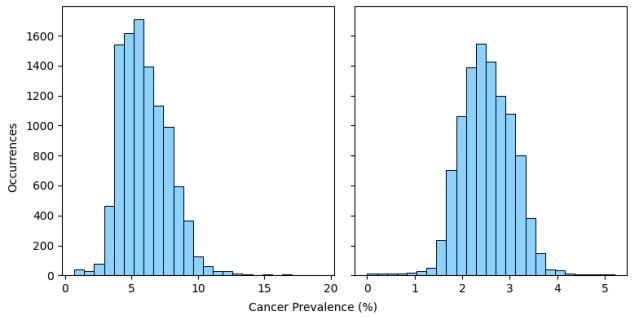

Fig. 2. Cancer prevalence pre- and post- Box-Cox transformation.

C. NCDs Overview

For the machine learning task, it is imperative to contextualize the output diseases, each of which holds unique implications towards the various facets of our data. These diseases were selected based on their prominence in existing society and need for further analysis.

Asthma is an, often chronic, respiratory disease caused by airway hyperresponsiveness, inflammation, and remodeling [14]. Its most common symptoms involve shortness of breath, particularly following periods of high physical activity. While it is uncommon for asthmatic symptoms to result in death, its impact towards individual quality-of-life is severe, resulting in sleep disturbances, physical limitations, detracted lung function, and use of prescription medications such as inhalers [15].

High blood pressure, or hypertension, involves the excess of cardiovascular function that stresses heart and blood vessel organs [16]. Particularly in the working-aged population, sustained levels of work- and social- induced stress have a prominent linkage towards hypertension contraction rates [17]. While medical treatments to diagnose and prevent the exacerbation of hypertension exists, untreated cases serve as a precursor to both short-term effects such as coronary heart disease and stroke [18] and long-term effects such as Alzheimer's Disease [19], both types of which degrade physical and emotional wellbeing and drastically increase associated mortality risks.

Cancer's prominence is marked by the current absence of an empirical cure. This disease is heavily influenced by individual lifestyles, most well-researched through excess ultraviolet exposure, which increases the risk of cancerous tumor mutation burdens irreparable by DNA damage response pathways [20], and the consumption of specific chemical- and bacteria- pervasive foods, such as red meats [21].

Diabetes is a hereditary disease contracted through sedentary lifestyles and overconsumption of glucose-rich foods [22] that holds a strong causal relationship with obesity [23]. This is caused by the autoimmune destruction of insulin-producing beta cells in the pancreas [24]. Existing treatments for diabetes include prescription medication, gene therapy, and lifestyle changes [25].

As one of the most rapidly proliferating global diseases, poor mental health is increasingly being studied within populations of all social classes. While poor mental health can be addressed through personalized therapeutic care, its surrounding stigmas often deter individuals from seeking help, risking longer-term consequences of suicide, alcoholism, and schizophrenia [26].

D. Model Overview

In total, eight regression models, selected on the basis of past predictive capabilities and applicability to this study's methodologies, were selected to evaluate each of the five NCDs. A brief overview of the selected models are described below, in alphabetical order:

1. Cubist: Based off of Quinlan's M5 model tree [27], this decision tree-based model utilizes a series of regressive models at the terminal leaves of the model. The parameters of these regressive models, which appear at every tree node, are dictated by the feature rules of the node.

2. Elastic Net: This linear regression addresses the individual sparse and regularization benefits of Lasso (L1) and Ridge (L2) regularization, respectively, by grouping features to either include or exclude in the resulting linear regression.

3. Gradient Boosting (GB): A sequential, tree-based regressor that aims to incrementally minimize the losses (residuals) of prior, weak decision trees. Every succeeding tree is fitted to the residuals of the past trees in order to build a better generalized model. For this regressor, learning rate is the most influential feature towards model outputs.

4. k-Nearest Neighbors (kNN): The k-Nearest Neighbors (kNN) algorithm, first proposed by Fix et al. [28], involved discriminately clustering a sample given an input sample of varying dimensions. In regressive tasks, the algorithm computes the weighted mean of a data point's K closest points, as dictated by the distance function of the algorithm.

5. Lasso: This L1 regularization model employs a feature-coefficient algorithm to create a sparse linear regression composed of a weighted subset of features. In doing so, the model effectively diminishes the effect of ineffectual features. Thus, it is commonly used as a tool for dimensionality reduction on high-complexity datasets.

6. Multi-Layered Perceptron (MLP): Developed by Rumelhart et al. [29], this supervised technique utilizes neurons within each successive layer to calculate portions of

the output. The resulting prediction is a summation of the individual outputs of every individual neuron in the final layer.

7. Random Forest (RF): The Random Forest model utilizes perturbation- the artificial addition of noise- to build a robust ensemble of weaker Decision Trees. In addition, its induction of randomness allows a more diverse set of individual trees to be built, addressing the variance limitations exhibited in said trees.

8. Ridge: This L2 regularization aims to impose penalties on certain feature coefficients, albeit not creating a sparse model. This shrinkage is particularly powerful towards addressing multicollinearity between features, effectively improving its linear regression performance.

E. Model Evaluation Metrics

Model evaluation is an imperative step in ensuring result validity. For this study, three evaluation metrics were used: coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

The equation for R^2 is shown above, where yi is the expected value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the mean of all expected values. This metric is particularly important in regression for its denotation of the correlation, or trend, between the predicted versus expected output. In cases where the error is sufficiently large, R^2 is able to achieve negative values, equivalent to a lack of any correlation between predicted and expected values.

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N}}$$

The equation for RMSE is shown above, where yi is the expected output, $\hat{y}_i$ is the predicted output, and N is the number of analyzed census tracts. This metric is particularly powerful due to its interpretability with its sensitivity to outliers, strictly penalizing the model for larger errors.

$$MAPE = \frac{\sum \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{N}$$

The equation for MAPE is shown above, where the summation of the difference of expected (y_i) versus predicted ($\hat{y}_i$) prevalence is averaged over N, the number of census tracts. This measure provides the average "accuracy", or predictive ability, of a model, and is comparable to related studies with varying output scales, due to its relative, rather than absolute, measure. This quantity is expressed on a flotational scale from 0 onwards, where a MAPE of 0 denotes no quantitative difference between the predicted and expected values.

III. Results

A. Model Results

Given the significance of hyperparameters towards influencing a machine learning model's underlying relationship extraction abilities, model hyperparameter tuning was conducted through a grid search. In this process, values were uniformly selected along a given range, with each value corresponding to its respective parameter. Given the large parameter search space for each of the eight models, a randomized grid search was used to conduct an evaluation on the largest range of possible parameter combinations. For this paper, the randomized search employed tested 100 different parameter combinations with 5-fold cross validation to ensure predictive robusticity. Following the grid, the top parameter combinations were selected from every model, their learning curves analyzed.

TABLE I. Top Model Results for All Outputs, Selected from MAPE

NCD	Model	R^2	RMSE	MAPE (%)
Asthma	Cubist	0.871	0.480	3.849%
HBP*	Cubist	0.803	2.186	6.077%
Cancer	Cubist	0.904	0.568	7.656%
Diabetes	Cubist	0.863	1.017	7.866%
PMH^	Cubist	0.904	0.953	5.531%

*(High Blood Pressure)
^(Poor Mental Health)

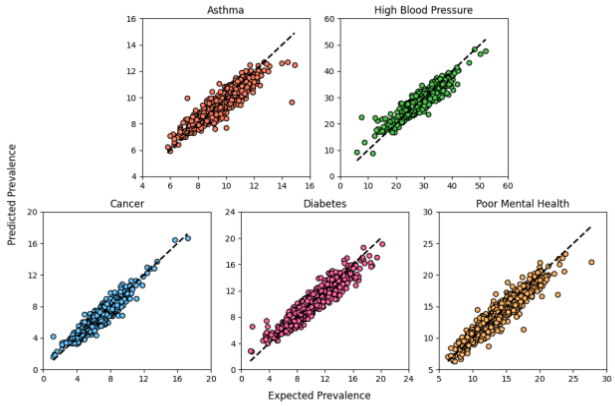

Fig. 3. Predicted versus expected prevalences for all five output NCDs.

While the models performed quite well, many of them contained overfitting between the training and validation curves. Thus, models were subsequently manually fine-tuned in order for their training and validation learning curves to reach a convergence threshold of 0.005 absolute difference in MAPE.

For every output, the Cubist, GB, MLP, and RF models all maintained significantly lower MAPE values as compared with the kNN, Lasso, Ridge, and Elastic Net models (Tables S2-6). Out of the four top performers, the Cubist model held the best performance across all three evaluation metrics throughout all outputs (Table I). Unsurprisingly, the MLP regression came second, achieving nearly negligible differences (≤~0.100% MAPE) compared with the Cubist model in predicting outputs such as the prevalence of Cancer and Poor Mental Health. For all top Cubist predictors, the R2 value maintained above 0.8 (0.86 excluding High Blood Pressure), demonstrating the high levels of explained variance carried by the models.

TABLE II. NCD Test Set and Validation Set Evaluation Discrepancies

	Test Set Results			|Performance Change|		
NCD	R^2	RMSE	MAPE	R^2	RMSE	MAPE
Asthma	0.857	0.510	3.903%	0.014	0.030	0.054%
HBP*	0.826	2.086	6.094%	0.023	0.100	0.017%
Cancer	0.915	0.545	7.690%	0.011	0.023	0.034%

Diabetes	0.886	0.909	7.684%	0.023	0.108	0.182%
PMH^	0.902	0.971	5.495%	0.002	0.018	0.036%

*(High Blood Pressure)
^(Poor Mental Health)

Following the cross-validated evaluation of the top Cubist models was performed, a test set evaluation was conducted, with the model's predicted versus actual values displayed (Fig. 3). Given the absence of perturbation in the Cubist algorithm, the entirety of the training set could be used to fit the model for prediction, ignoring the need for a cross validation step. The test set results were found to be extremely similar to validation results (Table II), with the highest absolute MAPE difference being only 0.182%. A geospatial comparison of the predicted versus expected tract values (Fig. 4) in the held-out test set affirms the accurate predictive capabilities of the Cubist model.

B. Feature Importances

In order to analyze the contribution of each feature in the models- and understand the driving factors behind each NCD- the feature importances were retrieved from every model. While the y-transformation for the predictions of Cancer required the use of a custom scoring function involving inverting the Box-Cox results using lambda, all other outputs required no alterations. An exhaustive permutation importances search was employed to determine the contribution of individual features. This algorithm randomly shuffles the values for every given input feature, recording the resulting model degradation; higher importance features will experience larger magnitudes of detracted model performance, and lower importance features may cause little difference in model performance, if at all. To ensure the robustness of this measure, every feature in the latter run cases was permuted 10 times, with the feature ranking derived from the mean increase in MAPE (Figs. S2-6). The top 15 features for every model were recorded and plotted (Fig. 5). In predicting Asthma and Poor Mental Health, the extent of poverty served unequivocally as the highest importance feature. When predicting High Blood Pressure and Cancer values, the percent of working-aged individuals served as the most important predictor. For the prediction of Diabetes, the percentage of minorities held the greatest significance, with the rate of poverty falling slightly behind.

The unpermuted importances from the Cancer model saw the percentage of working aged, under 17 years of age, and minorities as the top features, with other features falling close behind. Interestingly, despite empirical evidence between social and economic disparities linking proximity to pollutionary structures [30], their respective features generally contributed little to nothing for the model, ranking far lower than other social and demographic predictors. Still, out of these Environmental Justice factors, the percentage of a tract within impaired water bodies held the greatest significance for the prediction of Asthma and High Blood Pressure, and the percentage of houses predating 1980 held the greatest significance in High Blood Pressure, and Cancer, Diabetes, and Poor Mental Health prediction. For all models, the proportion of a tract within lead or coal mines held zero importance.

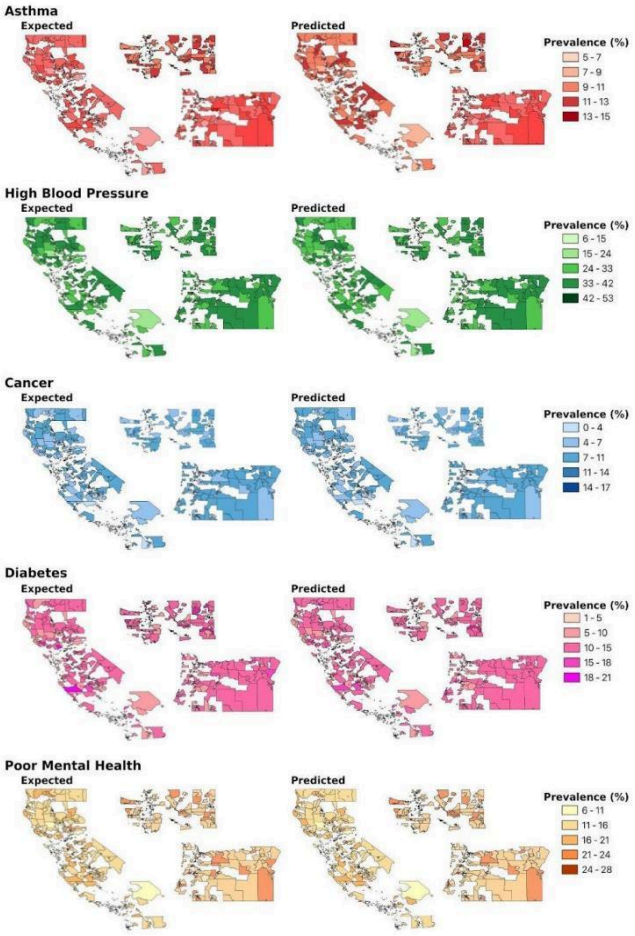

Fig. 4. Geospatial display of predicted versus expected prevalences, where California is shown left, Washington is shown upper right, and Oregon is shown lower right.

In order to further understand the relationship between various features and predictive power, the 5-fold cross-validation accuracy was plotted against the number of top features for every output (Fig. 6), with the top features starting from 5 and incrementing to 40, with a step size of 5 features. The said features were added sequentially from their order of importances. For every output, the most significant change in MAPE was seen between the addition of 5 to 10 features, with Asthma and Poor Mental Health seeing the lowest drops of ≤~1.2% MAPE, and High Blood Pressure and Diabetes seeing the highest drops of ≤~3% MAPE. While the top 15 sequentially selected features (in descending order of feature importance) were generally adequate enough to represent the model within 0.5% MAPE of the baseline, the top 20 features were required to achieve within 0.1% MAPE. Thus, all of the features beyond the 20th descending feature exhibited a negligible, or even negative, impact towards the model. To further support this claim, the NCD models (Table I) were retrained and retested using the subset of the top 20 features, achieving results comparable to the complete set of 41 features (Table III). Out of just the top 5 features, sociodemographic and employment determinants held the highest importance by far, with minority population, population under 200% poverty, and working-aged population

all appearing four times, population under 18 years of age appearing three times, and adult population without a secondary diploma appearing two times. Excluding proportion of tract inside coal- or lead-based structures, the bottom five features importances were all classed as Environment/Geographic, with the total land in a tract and proportion of tract area in an EPA-sanctioned Treatment, Storage, and Disposal site both appearing five times, proportion of tract area in an EPA-sanctioned National Priority List site and near an airport both appearing four times, and proportion of tract area in an EPA-sanctioned risk management plan site appearing three times.

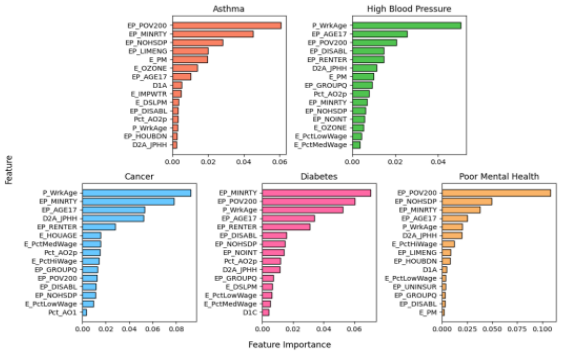

Fig. 5. Top 15 features using 10-repeat permutation importances.

In general, throughout all NCDs, the Sociodemographic factors pertaining to population under 18 and minorities held the highest importances, the Environment/Geographic factors pertaining to excess particulate matter 2.5 (PM2.5) concentration unequivocally held the highest importances, the Environmental Justice factors pertaining to proportion of a tract near an impaired water body and percentage of houses predating 1980 held the highest importances, and the Economic factors pertaining to tract-wide poverty levels and jobs per household holding the highest importances.

TABLE III. MODEL RESULTS WITH TOP 20 VALIDATION FEATURES

NCD	R^2	RMSE	MAPE (%)
Asthma	0.871	0.480	3.849%
HBP*	0.803	2.186	6.077%
Cancer	0.904	0.568	7.656%
Diabetes	0.863	1.017	7.866%
PMH^	0.904	0.953	5.531%

*(High Blood Pressure)
^(Poor Mental Health)

IV. DISCUSSIONS

A. Model Performance Analysis

Despite the regularized linear regressions and nearest neighbor algorithms all performing to an objectively adequately high degree, they are unable to capture the non-linear relationships as well as the top performing models. In particular, the k-Nearest Neighbors regressor, despite holding high relevance towards certain classification tasks [31], likely performed poorly on both training and testing data due to its dependence on high amounts of training relative to testing data, and further affected by the wide range of expected output prevalences.

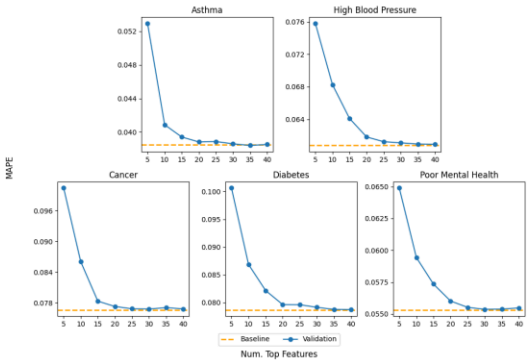

Fig. 6. Validation MAPE curves for a given number of top features.

For any reputable machine learning study, the goal is not only to conceive of an accurate model, but also understand or explain their underlying mechanisms to promote the greatest transparency between inputs and outputs. Traditionally, non-linear algorithms, such as neural networks or decision trees, have made use of random initialization to combat overfitting, and thus promote model reproducibility. In neural networks, for instance, the model is initialized with fixed weights and biases in the hidden layers in order to best adapt and "tune" these values as the model iterates, as opposed to manually setting weights and biases (ie. all to 0), which fails to promote similar accuracies [32]. Decision trees also employ randomization, specifically when selecting a subset of input features at nodes- their decision making step [33]. The random forest model, previously introduced as the aggregation of multiple decision trees, allows for the most robust and predictable models by increasing the number of randomized child predictors.

While randomization can be helpful in creating robust and generalized models, in addition to detecting extremely niche featural relationships, their varied component values cause the completed model to contain weariness surrounding objective feature importances. The Cubist model triumphs in this regard due to its use of deterministic, rather than probabilities, rules [30], producing the same endpoint linear regressions given the same dataset and hyperparameters, and thus, the same model results. This allows Cubist to be one of the top models in achieving explainable machine learning. Hence, other facets of the Cubist model require full understanding in order to best conclude on featural relationships. Given the non-linear relationships in the data, a single Cubist iteration was unable to capture the data's complexities. The boosting parameter, which was set to 30 for every single model, meant the model iterated thirty times over the cross validated data, using the same deterministic approach every iteration to revise and retune every rule. Rather surprisingly, during the tuning process, when altering the number of rules- the number of endpoint linear regressions- arbitrarily increasing the rule size both held negligible differences towards model accuracy and towards model overfitting, seen most prominently through the Asthma and Hypertension predictors employing 350 and 1006 rules, respectively; both nonetheless exhibited training and validation curves that converged within the threshold, attesting to the strong generalization abilities of the Cubist algorithm. Additionally, during the hyperparameter tuning process, it was found that the number of neighbors- the *k* nearest training

samples used to adjust a predicted output [34]- resulted in severe overfitting, creating training and validation learning curves with up to a 3% difference in MAPE.

When considering the diverse profiles of census tracts, primarily their varying levels of industrial development and qualities-of-life, the model results would be expected to contain multiple outliers or lower predictive accuracy. However, all outputs were predicted fairly well by the features. While the feature set still has room to improve, this standardized feature set marks a significant improvement from the works of Luo et al. [7] and Feng et al. [8]. Specifically, this paper analyzed disease prevalence on a census tract scale rather than state scale, providing far more granular results for hotspots of intervention; while a state may be shown to have a high prevalence of a certain NCD, it is infeasible to target intervention towards the entirety of the state. Rather, regional patterns in diseased populations help officials detect specific outliers, allowing them to more quickly address the driving causes of diseases. Additionally, the use of two standardized datasets, both of which were compiled through US government-sponsored programs, helps to promote consistency and unbiased data analysis within all geographic regions, as opposed to Feng et al.'s use of data from the 500 Cities Project and 311 service request data [8], both of which are available only for specific urban and suburban census tracts.

B. Disease Determinants

For all diseases, no tracts existed such that the disease was completely absent. Generally, diseases with shorter contraction periods and lower mortality rates held far higher prevalence; high blood pressure, poor mental health, diabetes, cancer, and asthma appeared in upwards of one half, one fourth, one fifth, one sixth, and one seventh of the population, respectively. Cancer, despite being the one outlier over asthma-present populations, reasonably holds the lowest average in tract prevalence. Out of all predictions, relatively few outliers persisted, with the largest concentration of outliers in predicting Poor Mental Health. In the feature sets used to predict said prevalence, sociodemographic factors by far held the highest importance, followed by economic, environmental justice, and environment and geographic features.

1) Generalized Lightweight Framework

For all NCDs, sociodemographic factors by far held the highest importance, followed by economic, environmental justice, and environment and geographic features. Additionally, throughout all outputs, the influence of tract proportion nearby coal and lead mines being 0 is reasonably attributed to their extremely high skew, as the structures are very rarely found throughout the West Coast, and hence, carry no predictive significance. However, had regions more prominently within these structures been trained and evaluated, a strong, positive correlation would be expected to be seen, primarily due to mine tailings and toxic particulate and chemical emissions [35].

Environmental factors, holding reasonably high importances within the top 20 features for Asthma, High Blood Pressure, and Cancer due to the relationship between said diseases and environmental influencers. In general, however, environmental and land use metrics held relatively low importances in this study due to inconsistent regional development metrics, establishing difficulty towards feature generalization.

Out of all sociodemographic factors, the percentage of minorities, working class individuals, and children exhibited the top correlation, in no particular order. Minority populations hold disproportionate disease influence as the result of discrimination; while overt discrimination towards minorities has long been outlawed in the United States, de facto factors play large roles in influencing the housing and employment opportunities provided to these groups, also resulting in increased exposure to harmful pollutants. These factors, classified as Environmental Justice factors, pertain to the proximity of a census tract to air pollution, noise pollution, water pollution, and radiation, where a strong correlation between minority and poverty rates and environmental injustice has been proven in seminal studies [36]; however, the models failed to find much importance in Environmental Justice features, given the lower quartile rankings of a majority of these structures and zones. The few structures that do hold significance, however, exhibit reasonably high feature importances. As the most significant structure, impaired watersheds are classified as watersheds holding excess amounts of contaminants harmful towards human bodies. Factors such as urban runoff resulting from impervious services introduces excess nutrients to nearby tributaries and watersheds, consequently causing eutrophication and undrinkable water [37]. Also, weaker regulatory standards, particularly in minority communities, results in disproportionate pollutant releases from industrial and household activity into nearby water bodies [38], resulting in similar, toxic exposure levels.

The substantial importance between poverty rates and minority population correlates with the conclusion of past literature. With the primary cause of NCDs being lifestyle choices, whether compulsory or self-decided, individuals with lower socioeconomic statuses will have increased contraction risks of all diseases. Select features of interest are outlined in the successive subsections.

2) Asthma analysis

Poorer communities, for one, lack the funding to afford adequate housing renovation measures, in addition to their aged houses, resulting in elevated concentrations of airborne compounds such as asbestos and structural detritus, both of which hold proven correlations with bronchitis and asthma contraction [39]. The environment/geographic features, while still holding objective high importance, ranked relatively lower than other economic and demographic determinants. Despite these factors- the mean days above regulatory PM2.5 concentrations, mean days above regulatory Ozone concentrations, and ambient diesel concentration- all directly affecting asthmatic progression, sociodemographic factors likely hold greater importance due to their greater influence towards rendering communities unable to address asthma sources. Similar to houses, unsafe employment conditions, holding a significant association with respiratory irritants, disproportionately affect minority, undereducated, and disabled populations [40].

3) High Blood pressure analysis

For High Blood Pressure, the working aged population held the highest correlation. This disease, unlike other diseases, is relatively similar across socioeconomic status, with slightly higher contraction rates found in low- and mid-income as opposed to high-income regions [41]. The working-aged population, the most important factor, of any society experiences great levels of work-induced stress, consequently resulting in the adopting adverse behaviors that increase hypertension contraction rates. The second highest factor being the population under 18 is explained through physical and emotional stressors, as adolescents have been found to be the age group most adversely affected by stress [42]. Their stress is primarily found through school and family responsibilities, coupled with their inherent quicker adoption of unhealthy behaviors and poorer coping mechanisms relative to adults [43].

4) Cancer Analysis

In Cancer predictions, the working aged population is the most significant factor due to work related exposure [44]. Additionally, in the context of environmental justice, minorities, the second most important feature, traditionally face blue-collar working conditions and housing opportunities with the highest level of carcinogenic exposure [45]. This is supported by the high importance of the housing age variable, confirming speculation that some communities may not have the adequate resources to both evaluate housing quality and implement the renovations required to mitigate the presence of pollutants such as asbestos, lead contamination, and radon poisoning from underground storage compartments [46]. Surprisingly, the proximity to impaired water bodies containing carcinogens is relatively low, attesting that many watersheds have likely been tested thoroughly for water contamination prior to consumption.

5) Diabetes Analysis

In a multitude of research papers, Diabetes is by far most heavily influenced by minority populations. This fact can be attributed to the racially-influenced construction of alcohol outlets within minority communities [47]. This idea could be further supported through poverty rates, as these shops may mark their products cheaper than less-accessible, traditional markets. Other causes, such as white-collar office roles, causes individuals to adopt sedentary lifestyles, and, in conjunction with high blood pressure, often leads to elevated contraction risks for diabetes [48]. Interestingly, the inclusion of households with more than two cars holds reasonable significance, and is supported by existing research to embody a negative correlation due to improved lifestyle habits as economic stability improves [49].

6) Poor Mental Health Analysis

By far, the most important predictor of Poor Mental Health is poverty. A linkage between poverty, lower self-esteem, increased psychological stress, and detracted mental health has been well-researched [50]. Despite common understanding of the work, financial, and familial responsibilities of adults, the population of children held surprisingly high importance within the model. Additionally, there exists a prominent stigma surrounding therapy and seeking professional help, particularly within adolescents [51], leading many early-stage cases of poor mental health and high blood pressure to proliferate.

C. Limitations and Future Work

Although our results pose significant implications towards the future of public care, self care, and policy, the NCD prevalence may be subject to varying degrees of reporting biases, particularly among households without convenient access to medical providers to report their health. In addition, census tracts with smaller populations means that every contracted case of an NCD would hold greater proportion than larger tracts, possibly misrepresenting the data and causing some discrepancies between predicted versus expected results. Despite these limitations, the model predictions still remained reasonably robust.

Future work regarding our results include using more specific featural breakdowns, such as various poverty and economic levels rather than their aggregation into binary classes (ie. below 200% poverty or not). This not only helps the model learn even more specific features, potentially increasing model predictive capabilities, but also allows for greater transparency in understanding granular features. Other features of interest include past NCD metrics, such past diagnoses, which could provide insight into disease duration and trends in contractions.

V. Conclusion

In this paper, the presence of five NCDs, Asthma, High Blood Pressure, Cancer, Diabetes, and Poor Mental Health was explored and analyzed. Using standardized databases collected by the US EPA and CDC, separated into census tracts in the West Coast states California, Oregon, and Washington, the Cubist model topped the predictive accuracy charts, outcompeting other top non-linear models. Through the machine learning approach presented in this paper, all five of the presented diseases are proven to be both predictable, interpretable, and reliable, unlocking the potential for researchers to provide targeted intervention schemes towards the most vulnerable communities nationwide.

Supplemental Materials

The supporting information can be downloaded here.

References

[1] Piovani, D., Nikolopoulos, G. K., & Bonovas, S. Non-communicable diseases: the invisible epidemic. Journal of Clinical Medicine, 2022, 11(19), 5939. [Google Scholar]

[2] World Health Organization. Noncommunicable diseases. World Health Organization (2023, September 16). https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[3] Samuel, A. L. Some studies in machine learning using the game of checkers. IBM Journal of research and development, 1959, 3(3), 210-229. [Google Scholar]

[4] Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 2021, 191, 487-492. [Google Scholar]

[5] Stephen, O., Sain, M., Maduh, U. J., & Jeong, D. U. An efficient deep learning approach to pneumonia classification in healthcare. Journal of healthcare engineering, 2019(1), 4180949. [Google Scholar]

[6] Bertini, A., Salas, R., Chabert, S., Sobrevia, L., & Pardo, F. Using machine learning to predict complications in pregnancy: a systematic review. Frontiers in bioengineering and biotechnology, 2022, 9, 780389. [Google Scholar]

[7] Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., Gupta, S., ... & Allender, S. Is demography destiny? Application of machine learning

techniques to accurately predict population health outcomes from a minimal demographic dataset. PloS one, 2015, 10(5), e0125602. [Google Scholar]

[8] Feng, C., & Jiao, J. Predicting and mapping neighborhood-scale health outcomes: A machine learning approach. Computers, Environment and Urban Systems, 2021, 85, 101562. [Google Scholar]

[9] Felix, E. A., & Lee, S. P. Systematic literature review of preprocessing techniques for imbalanced data. Iet Software, 2019, 13(6), 479-496. [Google Scholar]

[10] Yeo, I. K., & Johnson, R. A. A new family of power transformations to improve normality or symmetry. Biometrika, 2000, 87(4), 954-959. [Google Scholar]

[11] Branco, P., Torgo, L., & Ribeiro, R. P. SMOGN: a pre-processing approach for imbalanced regression. In First international workshop on learning with imbalanced domains: Theory and applications (pp. 36-50). PMLR, 2017 (October). [Google Scholar]

[12] Box, G. E., & Cox, D. R. An analysis of transformations. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1964, 26(2), 211-243. [Google Scholar]

[13] Osborne, J. Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research, and Evaluation, 2010, 15(1). [Google Scholar]

[14] Zuyderduyn, S., Sukkar, M. B., Fust, A., Dhaliwal, S., & Burgess, J. K. Treating asthma means treating airway smooth muscle cells. European Respiratory Journal, 2008, 32(2), 265-274. [Google Scholar]

[15] Globe, G., Martin, M., Schatz, M., Wiklund, I., Lin, J., von Maltzahn, R., & Mattera, M. S. Symptoms and markers of symptom severity in asthma—content validity of the asthma symptom diary. Health and quality of life outcomes, 2015, 13, 1-9. [Google Scholar]

[16] Heagerty, A. M., Heerkens, E. H., & Izzard, A. S. Small artery structure and function in hypertension. Journal of cellular and molecular medicine, 2010, 14(5), 1037-1043. [Google Scholar]

[17] Liu, M. Y., Li, N., Li, W. A., & Khan, H. Association between psychosocial stress and hypertension: a systematic review and meta-analysis. Neurological research, 2017, 39(6), 573-580. [Google Scholar]

[18] Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. The Lancet, 2002, 360(9349), 1903-1913. [Google Scholar]

[19] Joas, E., Bäckman, K., Gustafson, D., Östling, S., Waern, M., Guo, X., & Skoog, I. Blood pressure trajectories from midlife to late life in relation to dementia in women followed for 37 years. Hypertension, 2012, 59(4), 796-801. [Google Scholar]

[20] Lee, J. W., Ratnakumar, K., Hung, K. F., Rokunohe, D., & Kawasumi, M. Deciphering UV‐induced DNA damage responses to prevent and treat skin cancer. Photochemistry and photobiology, 2020, 96(3), 478-499. [Google Scholar]

[21] Chan, D. S., Lau, R., Aune, D., Vieira, R., Greenwood, D. C., Kampman, E., & Norat, T. Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. PloS one, 2011, 6(6), e20456. [Google Scholar]

[22] Wu, Y., Ding, Y., Tanaka, Y., & Zhang, W. Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. International journal of medical sciences, 2014, 11(11), 1185. [Google Scholar]

[23] Iglay, K., Hannachi, H., Joseph Howie, P., Xu, J., Li, X., Engel, S. S., ... & Rajpathak, S. Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus. Current medical research and opinion, 2016, 32(7), 1243-1252. [Google Scholar]

[24] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes care, 2010, 33(Supplement_1), S62-S69. [Google Scholar]

[25] Aloke, C., Egwu, C. O., Aja, P. M., Obasi, N. A., Chukwu, J., Akumadu, B. O., ... & Achilonu, I. Current advances in the management of diabetes mellitus. Biomedicines, 2022, 10(10), 2436. [Google Scholar]

[26] Inskip, H., Harris, C., & Barraclough, B. Lifetime risk of suicide for affective disorder, alcoholism and schizophrenia. The British Journal of Psychiatry, 1998, 172(1), 35-37. [Google Scholar]

[27] Quinlan, J. R. Learning with continuous classes. In 5th Australian joint conference on artificial intelligence, 1992 (November), Vol. 92, pp. 343-348. [Google Scholar]

[28] Fix, E. Discriminatory analysis: nonparametric discrimination, consistency properties (Vol. 1). USAF school of Aviation Medicine. 1985. [Google Scholar]

[29] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. Learning representations by back-propagating errors. nature, 1986, 323(6088), 533-536. [Google Scholar]

[30] Hajat, A., Hsia, C., & O'Neill, M. S. Socioeconomic disparities and air pollution exposure: a global review. Current environmental health reports, 2015, 2, 440-450. [Google Scholar]

[31] Taunk, K., De, S., Verma, S., & Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. In 2019 international conference on intelligent computing and control systems (ICCS) (pp. 1255-1260). IEEE. 2019 (May). [Google Scholar]

[32] Suganthan, P. N., & Katuwal, R. On the origins of randomization-based feedforward neural networks. Applied Soft Computing, 2021, 105, 107239. [Google Scholar]

[33] Dhurandhar, A., & Dobra, A. Probabilistic Characterization of Random Decision Trees. Journal of Machine Learning Research, 2008, 9(10). [Google Scholar]

[34] Kuhn, M., & Johnson, K. Applied predictive modeling (Vol. 26, p. 13). New York: Springer. 2013. [Google Scholar]

[35] León-Mejía, G., Vargas, J. E., Quintana-Sosa, M., Rueda, R. A., Pérez, J. P., Miranda-Guevara, A., ... & Henriques, J. A. P. Exposure to coal mining can lead to imbalanced levels of inorganic elements and DNA damage in individuals living near open-pit mining sites. Environmental Research, 2023, 227, 115773. [Google Scholar]

[36] Brodkin, C. A., Barnhart, S., Anderson, G., Checkoway, H., Omenn, G. S., & Rosenstock, L. Correlation between respiratory symptoms and pulmonary function in asbestos-exposed workers. American Review of Respiratory Disease, 1993, 148, 32-32. [Google Scholar]

[37] Li, L., Yu, Q., Gao, L., Yu, B., & Lu, Z. The effect of urban land-use change on runoff water quality: a case study in Hangzhou City. International Journal of Environmental Research and Public Health, 2021, 18(20), 10748. [Google Scholar]

[38] Bell, M. L., & Ebisu, K. Environmental inequality in exposures to airborne particulate matter components in the United States. Environmental health perspectives, 2012, 120(12), 1699-1704. [Google Scholar]

[39] Singh, A., Avula, A., & Zahn, E. Acute bronchitis. 2017. [Google Scholar]

[40] Burgard, S. A., & Lin, K. Y. Bad jobs, bad health? How work and working conditions contribute to health disparities. American Behavioral Scientist, 2013, 57(8), 1105-1127. [Google Scholar]

[41] Mills, K. T., Bundy, J. D., Kelly, T. N., Reed, J. E., Kearney, P. M., Reynolds, K., ... & He, J. Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. Circulation, 2016, 134(6), 441-450. [Google Scholar]

[42] Nagabharana, T. K., Joseph, S., Rizwana, A., Krishna, M., Barker, M., Fall, C., ... & Krishnaveni, G. V. What stresses adolescents? A qualitative study on perceptions of stress, stressors and coping mechanisms among urban adolescents in India. Wellcome open research, 2021,6. [Google Scholar]

[43] Garnefski, N., Legerstee, J., Kraaij, V., van Den Kommer, T., & Teerds, J. A. N. Cognitive coping strategies and symptoms of depression and anxiety: A comparison between adolescents and adults. Journal of adolescence, 2002, 25(6), 603-611. [Google Scholar]

[44] Yang, T., Qiao, Y., Xiang, S., Li, W., Gan, Y., & Chen, Y. Work stress and the risk of cancer: a meta‐analysis of observational studies. International Journal of Cancer, 2019, 144(10), 2390-2400. [Google Scholar]

[45] Lloyd, J. W. Long-term mortality study of steelworkers V. Respiratory Cancer in coke plant workers. Journal of Occupational Medicine, 1971, 13(2), 53-68. [Google Scholar]

[46] Yazzie, S. A., Davis, S., Seixas, N., & Yost, M. G. Assessing the impact of housing features and environmental factors on home indoor radon concentration levels on the Navajo nation. International journal of environmental research and public health, 2020, 17(8), 2813. [Google Scholar]

[47] Lee, J. P., Ponicki, W., Mair, C., Gruenewald, P., & Ghanem, L. What explains the concentration of off-premise alcohol outlets in Black neighborhoods?. SSM-Population Health, 2020, 12, 100669. [Google Scholar]

[48] Sitorus, N., Hanafi, A. S., & Simbolon, D. Joint effect of high blood pressure and physical inactive on diabetes mellitus: a population-based cross-sectional survey. Journal of Preventive Medicine and Hygiene, 2020, 61(4), E614. [Google Scholar]

[49] Bird, Y., Lemstra, M., Rogers, M., & Moraros, J. The relationship between socioeconomic status/income and prevalence of diabetes and associated conditions: A cross-sectional population-based study in Saskatchewan, Canada. International journal for equity in health, 2015, 14, 1-8. [Google Scholar]

[50] Knifton, L., & Inglis, G. Poverty and mental health: policy, practice and research implications. BJPsych bulletin, 2020, 44(5), 193-196. [Google Scholar]

[51] Pedersen, E. R., & Paves, A. P. Comparing perceived public stigma and personal stigma of mental health treatment seeking in a young adult sample. Psychiatry research, 2014, 219(1), 143-150. [Google Scholar]