# Cardiovascular Disease Risk Prediction via Social Media

**Al Zadid Sultan Bin Habib, Md Asif Bin Syed, Md Tanvirul Islam, Donald A. Adjeroh**
West Virginia University, Morgantown, WV 26506, USA
Email: {ah00069, ms00110, mi00018}@mix.wvu.edu, donald.adjeroh@mail.wvu.edu

SBP-BRiMS 2023

## Introduction

- The research underscores Cardiovascular Disease (CVD) risks via emotions expressed on social media, especially in the Appalachian region of the US.
- How psychological traits and social media data contribute to understanding CVD risk factors.
- Centers on using Twitter to monitor public discussions about CVD and employing NLP for sentiment analysis.
- Hybrid CNN-LSTM approach for predicting CVD risk and mentions of using ML algorithms.
- Focuses on creating an NLP dictionary for CVD using relevant keywords.
- Describes the data collection process, including state selection and using CDC datasets for demographics.
- Highlights Twitter data's potential for aiding public health practitioners in predicting CVD.

## Goals

- Develop a novel CVD-related keyword dictionary through sentiment analysis of Twitter data and employ ML models to predict CVD risk.
- Establish the relationship between psychological characteristics and risks of CVDs.
- Develop and evaluate a predictive model.

## Novelty

- NLP-based research is still less explored
- New Dictionary for CVD Risk prediction
- Twitter dataset vs CDC dataset comparison

## Acknowledgement

- Work supported in part by grants from the US National Science Foundation (Award #1920920, #2125872)

## Key References

- Eichstaedt, J.C., et al.: Psychological language on twitter predicts county-level heart disease mortality. Psychological science 26(2), 159–169 (2015)
- Sinnenberg, L., et al.: Twitter as a potential data source for cardiovascular disease research. JAMA cardiology 1(9), 1032–1036 (2016)



## Developed Dictionary

- Psychological Keywords
- Clinical Keywords
- anesthesia, angiogram, cardiologist, echocardiogram, heart attack, heart failure, hypertension, chest pain, stress, smoking, cholesterol and alcohol use

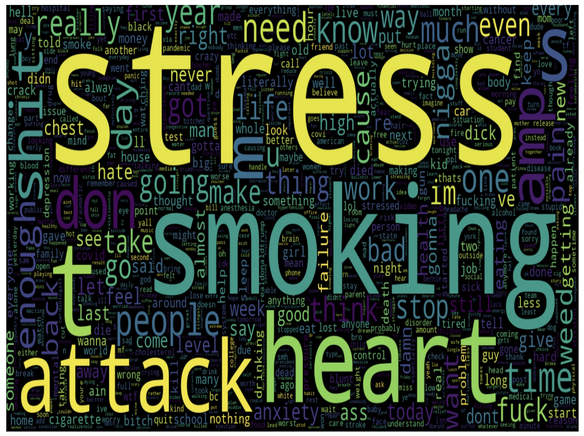


Fig. 1: Generated Word Cloud for the Tweets.

## Selected States

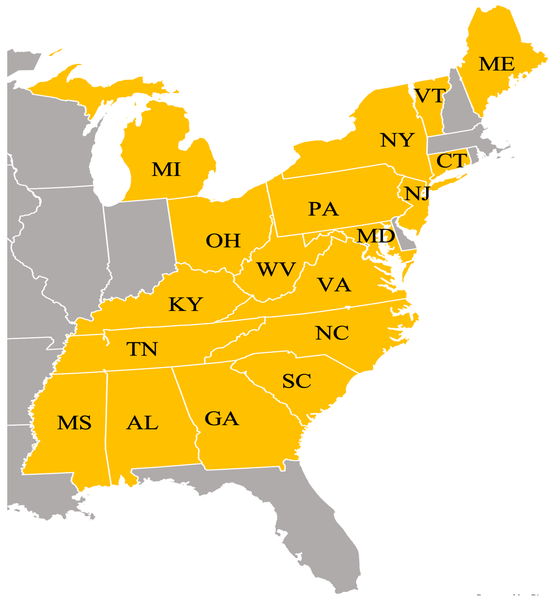


Fig. 2: Selected States to Collect Tweets.

## Overall Workflow

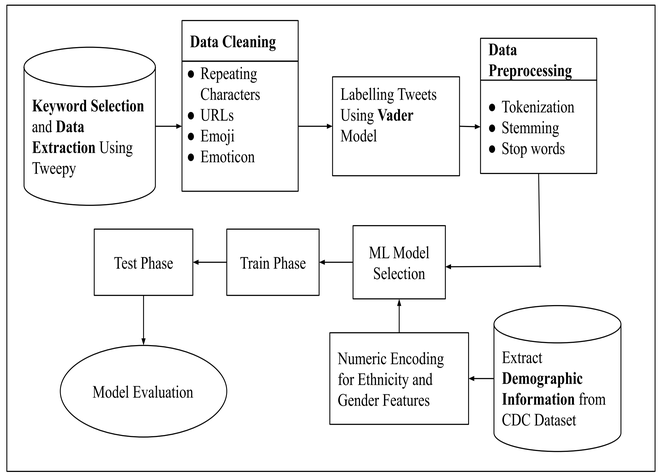


Fig. 3: Overall Workflow for Proposed Framework.

## Class Distribution and Architecture

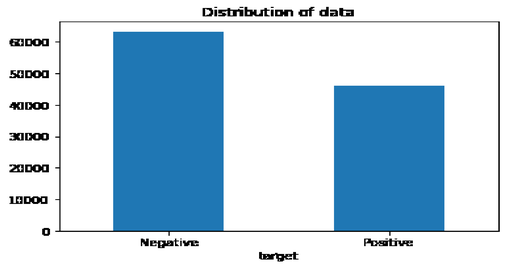


Fig. 4: Twitter Data Class Distribution.

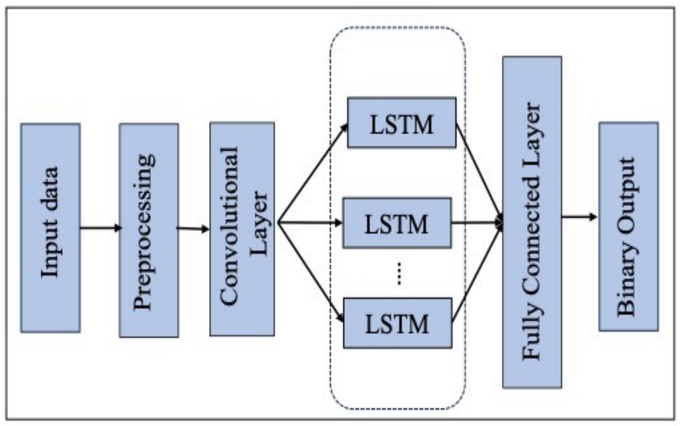


Fig. 5: CNN-LSTM Hybrid Architecture.

## Results Analysis

**Table 1.** Performance Evaluation for the Twitter Based Dataset.

| Model | Test Accuracy | Precision | Recall | F1 | MCC | CK |
|---|---|---|---|---|---|---|
| CNN-LSTM | 77.51% | 0.75 | 0.68 | 0.72 | 0.53 | 0.53 |
| BNB | 74.55% | 0.84 | 0.48 | 0.61 | 0.48 | 0.44 |
| SVM | 88.75% | 0.87 | 0.86 | 0.86 | 0.77 | 0.77 |
| LR | 87.82% | 0.85 | 0.86 | 0.85 | 0.75 | 0.75 |
| CatBoost | 76.67% | 0.73 | 0.71 | 0.72 | 0.53 | 0.53 |

**Table 2.** Performance Evaluation for the CDC Dataset.

| Model | Test Accuracy | Precision | Recall | F1 | MCC | CK |
|---|---|---|---|---|---|---|
| CNN-LSTM | 57.64% | 0.63 | 0.36 | 0.45 | 0.17 | 0.15 |
| BNB | 57.93% | 0.67 | 0.31 | 0.42 | 0.19 | 0.16 |
| SVM | 57.55% | 0.61 | 0.41 | 0.49 | 0.16 | 0.15 |
| LR | 58.03% | 0.62 | 0.39 | 0.48 | 0.17 | 0.16 |
| CatBoost | 57.42% | 0.61 | 0.42 | 0.50 | 0.16 | 0.15 |

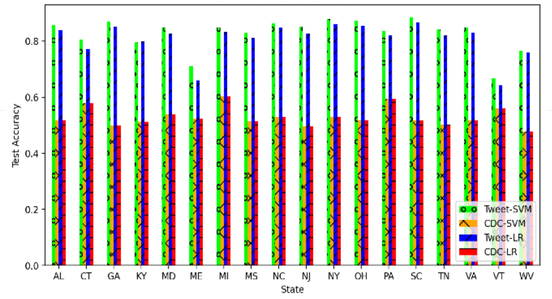


Fig. 6: Test Accuracy for SVM and LR.

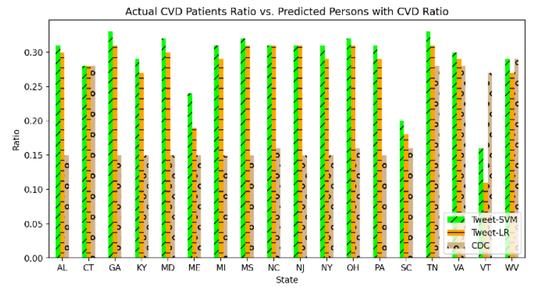


Fig. 7: Actual CVD Patients Ratio vs Predicted Persons with CVD Ratio.

## CONCLUSIONS

This research used sentiment analysis and ML to predict CVD risk from tweets in 18 US states, with SVM achieving the highest accuracy of 88.75%. Twitter data outperformed demographic information, highlighting its potential for CVD risk prediction.