MACHINE LEARNING MODEL FOR COVID-19 DIAGNOSIS

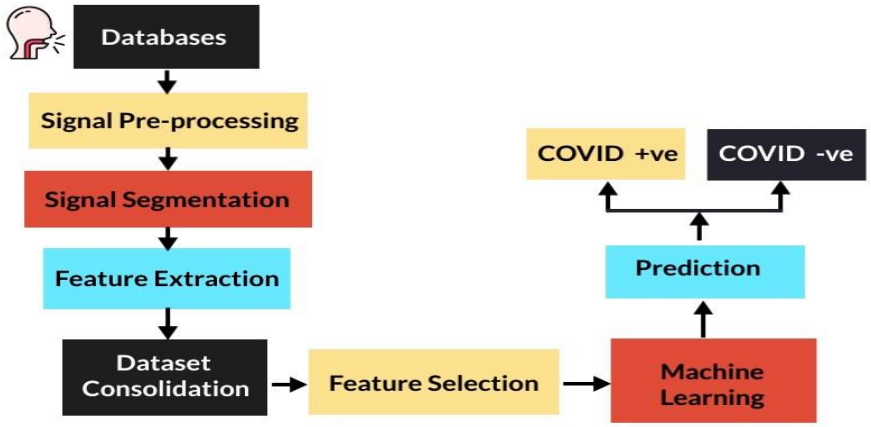
INTRODUCTION

In response to the global need for accessible and efficient COVID-19 diagnosis, this project focuses on utilizing cough sounds for early detection. With over millions of confirmed cases reported by the World Health Organization, there's a critical demand for non-invasive and cost-effective diagnostic methods. By extracting relevant features from cough sequences and employing machine learning techniques, this project aims to develop a simple yet effective model for COVID-19 diagnosis. The innovative approach not only showcases my expertise in data preprocessing and feature extraction but also highlights the potential of data analysis in addressing real-world healthcare challenges.

DATASET OVERVIEW: VIRUFY COVID-19 COUGH DATABASE

The project utilized the open-source VIRUFY COVID-19 cough database for classifying COVID-positive and negative coughs. It contained annotated data based on RT-PCR test results, along with patient demographics. With 121 pre-processed, segmented coughs from 16 patients, the dataset provided essential insights for the project, including original recordings and pre-processed cough samples in separate folders.

METHODOLOGY



DATA PREPROCESSING OF COUGH SOUND ANALYSIS

To prepare the audio data from the VIRUFY database for machine learning models, several pre-processing steps were implemented using specialized tools. De-noising with DWT, MATLAB's Wavelet Denoiser Application applied Discrete Wavelet Transform using Symlets wavelets for noise removal. Signal Enhancement using NCH suite Wavepad normalized and amplified audio signals for improved feature extraction. Signal Segmentation Using Wavepad, cough sounds were isolated from background noise, resulting in 193 cough samples for analysis.

FEATURE EXTRACTION:

The feature extraction process involved using MATLAB's Audio Feature Extractor to extract approximately 25 features from audio data, including MFCC, GTCC, zero-crossing rate, and

spectral features. Additionally, Jitter and Shimmer features were extracted using Praat software. These features were crucial for comprehensively analyzing cough sounds.The dataset, consolidated with extracted features and labels from RT-PCR test results, comprised a feature vector of size 193x125.

The extracted features include spectral domain features such as MFCC (Mel-Frequency Cepstral Coefficients), GTCC (Gammatone Cepstral Coefficients), harmonic ratio, spectral centroid, spectral kurtosis, spectral skewness, spectral spread, spectral entropy, spectral flux, spectral roll-off point, spectral crest, spectral decrease, and spectral flatness, as well as spectral flow. Additionally, temporal domain features like zero-crossing rate (ZCR), jitter, and shimmer were extracted from cough segments. These features collectively provide comprehensive insights for further analysis and classification tasks.

FEATURE SELECTION:

The feature selection process was conducted using various techniques implemented in Python libraries. Firstly, the Minimum Redundancy Maximum Relevance (MRMR) approach was applied using the `mrmr_selection` package, which identifies features with high relevance to the target variable (COVID-positive or COVID-negative coughs) while minimizing redundancy among selected features. Additionally, the ANOVA F-test from the `sklearn.feature_selection` module was employed to assess the statistical significance of features in discriminating between the two classes. Mutual information-based feature selection was performed using the `mutual_info_classif` function from the same module, which evaluates the dependency between features and the target variable. These methods helped in identifying a subset of features with the most discriminatory power, including coefficients from GTCC and MFCC, as well as spectral and temporal domain features such as spectral skewness, flux, entropy, slope, spread, zero crossing rate, jitter, and shimmer. The selected features were then used for further analysis and classification tasks to differentiate between COVID-positive and COVID-negative cough segments.

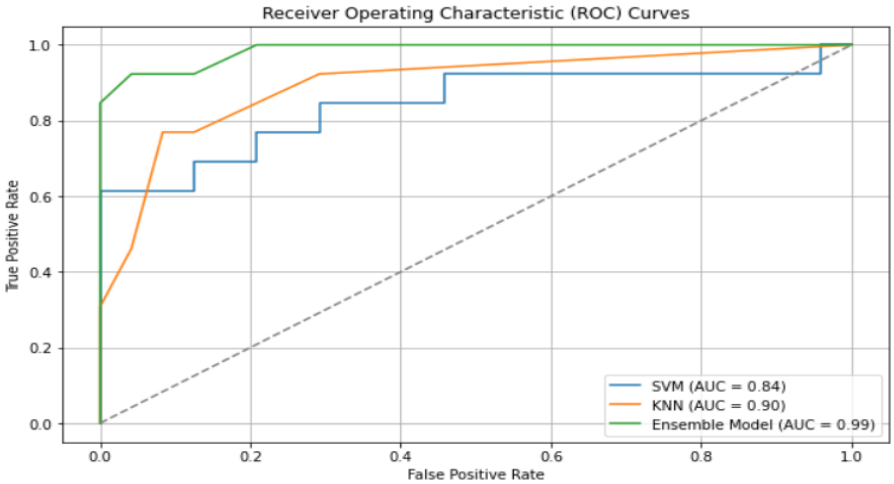
MODEL TRAINING

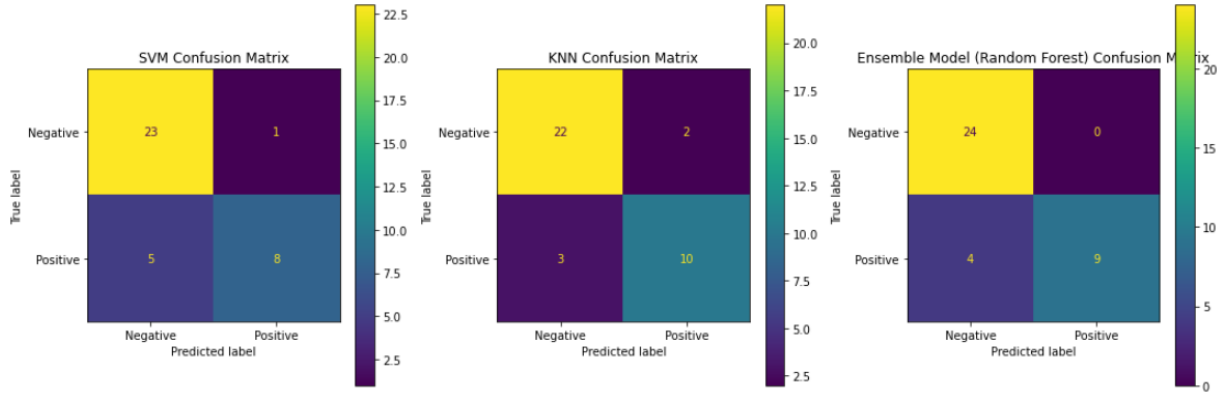
The dataset was split into training and testing sets, and StandardScaler was used for feature scaling. Three models were trained and evaluated: K-Nearest Neighbors (KNN) with k=5, Random Forest with 100 trees, and Support Vector Machine (SVM). All models were assessed using the testing set. KNN achieved an accuracy of 86.49%, Random Forest achieved 89.19% accuracy, and SVM achieved 84.51% precision and 83.78% recall. This rigorous evaluation helped select the most appropriate model for the dataset.

MODEL EVALUATION

The evaluation of the three models - Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and an Ensemble Model using Random Forest - revealed the Ensemble Model to be the superior choice for this dataset. The Receiver Operating Characteristic (ROC) curves showed that the Ensemble Model had the highest Area Under the Curve (AUC) of 0.99, indicating excellent predictive ability, while KNN and SVM had lower AUCs of 0.90 and 0.84,

respectively. Furthermore, the Ensemble Model's confusion matrix demonstrated its strength, with only 4 instances misclassified as false negatives, compared to 5 misclassifications for KNN (3 false negatives, 2 false positives) and 6 for SVM (5 false positives, 1 false negative). Quantitative evaluation metrics further solidified the Ensemble Model's dominance, with the highest precision (0.9073), recall (0.8919), and F1-score (0.8862), followed by KNN (precision: 0.8636, recall: 0.8649, F1-score: 0.8635) and SVM (precision: 0.8451, recall: 0.8378, F1-score: 0.8293). Collectively, these results conclusively establish the Ensemble Model as the most accurate and reliable choice for classifying this particular dataset.





CONCLUSION:

In this project, a multifaceted exploration of data analysis and machine learning was undertaken, with a particular focus on COVID-19 diagnosis using cough sound. Through meticulous data preprocessing and feature selection, the dataset was curated to prepare it for modeling. Leveraging Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and a Random Forest Ensemble Model, each model's performance was rigorously evaluated. Notably, the findings highlighted the Ensemble Model's superiority in accurately diagnosing COVID-19 based on cough sound, showcasing its remarkable predictive capabilities. This endeavor underscores the transformative potential of advanced data analysis and machine learning techniques

revolutionizing healthcare practices, particularly in the context of infectious disease detection and diagnosis.