****

**A Machine Learning Approach of Risk Assessment for Contagious Disease Like COVID-19 by Analyzing Lifestyle**

**Project Thesis**

**Submitted By**

|  |  |
| --- | --- |
| 17-33438-1 | Siam, Nafiz Fuad |
| 17-33365-1 | Khan, Mahira Tabassum |
| 17-33376-1 | Rownak, M. R |
| 17-33420-1 | Juel, Md. Rejaben Jamin |

**Department of Computer Science**

**Faculty of Science & Technology**

**American International University-Bangladesh**

**August 2021**

|  |
| --- |
| Declaration |

We declare that this thesis is our original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Siam, Nafiz Fuad**  17-33438-1  Computer Science and Engineering |  | **Khan, Mahira Tabassum**  17-33365-1  Computer Science and Engineering |
|  |  |  |
|  |  |  |
| **Rownak, M. R**  17-33376-1  Computer Science and Engineering |  | **Juel, Md. Rejaben Jamin**  17-33420-1  Computer Science and Engineering |

|  |
| --- |
| Approval |

The thesis titled "A Machine Learning Approach of Risk Assessment for Contagious Disease Like COVID-19 by Lifestyle Analysis" has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science on (date of defence) and has been accepted as satisfactory.

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Dr. Ashraf Uddin**  Assistant Professor & Supervisor  Department of Computer Science  American International University-Bangladesh | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Dr. Dip Nandi**  Associate Professor & Director & External  Faculty of Science & Technology  American International University-Bangladesh |
|  |  |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Dr. Md. Mahbub Chowdhury Mishu**  Assistant Professor & Head (Undergraduate)  Department of Computer Science  American International University-Bangladesh | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Professor Dr. Tafazzal Hossain**  Dean  Faculty of Science & Information Technology  American International University-Bangladesh |
|  |  |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Dr. Carmen Z. Lamagna**  Vice Chancellor  American International University-Bangladesh | |

|  |
| --- |
| Acknowledgment |

Let us begin by expressing our thankfulness towards Allah, the Almighty, for enabling us to complete this thesis on schedule.

We would love to show our heartfelt gratitude to our respected supervisor Dr. Ashraf Uddin Sir, Assistant Professor, from the bottom of our hearts for his unwavering motivation, critical advice, supervision, instructions, guidance, and recommendations, as well as for consistently encouraging us and supporting us through his hands and pens to complete our project thesis without any hassle at American International University-Bangladesh (AIUB).

Furthermore, it would be our pleasure to thank our external Dr. Dip Nandi Sir, Assistant Professor and Director, Faculty of Science and Technology, for his useful suggestions and recommendations, professional directives, and compassionate support in updating our work throughout this period.

Thanks to everyone on our department's personnel and the AIUB administration for their assistance and cooperation throughout this study.

|  |
| --- |
| Abstract |

During the coronavirus pandemic, people were advised to maintain a regulated and suggested hygienic lifestyle to prevent mass transmission. This study has used machine learning to analyze lifestyle-specific data to see the relevancy of such claims and to prepare a machine learning model/system that can predict whether a person is affected or going to be affected by the disease. A public survey was done on lifestyle-related questions that resulted in a dataset consisting of 620 responses. Typical machine learning methodology has been followed that contains steps like data preprocessing, feature engineering, training, evaluation, etc. An iterative strategy has been used during the training and evaluation phase of the study. This study has used three machine-learning algorithms; one of them is a Neural Network. Relevancy of hygienic lifestyle habits in case of preventing such disease has been found. The study has developed a system by evaluating and selecting a machine learning model that can predict if a person is affected or going to be affected by such a disease. At the end of this study, limitations and scope for future improvements have also been discussed.

|  |
| --- |
| Table of Contents |

[Declaration 2](#_Toc79951857)

[Approval 3](#_Toc79951858)

[Acknowledgment 4](#_Toc79951859)

[Abstract 5](#_Toc79951860)

[Table of Contents 6](#_Toc79951861)

[List of Tables 8](#_Toc79951862)

[List of Figures 9](#_Toc79951863)

[Chapter 1: Introduction 10](#_Toc79951864)

[1.1 The motivation of the study 11](#_Toc79951865)

[1.2 Symptom Analysis 11](#_Toc79951866)

[1.3 Lifestyle Analysis 12](#_Toc79951867)

[Chapter 2: Literature Review 14](#_Toc79951868)

[Chapter 3: Methodology 22](#_Toc79951869)

[3.1 Dataset Description 22](#_Toc79951870)

[3.2 Machine Learning Approach 31](#_Toc79951871)

[3.2.1 Data Preparation 32](#_Toc79951872)

[3.2.1.1 Column Renaming 32](#_Toc79951873)

[3.2.1.2 Extraction of Columns 33](#_Toc79951874)

[3.2.1.3 One-Hot Encoding 33](#_Toc79951875)

[3.2.1.4 Scaling Features 35](#_Toc79951876)

[3.2.1.5 Balancing Dataset 35](#_Toc79951877)

[3.2.2 Training and Evaluation 37](#_Toc79951878)

[3.2.2.1 Algorithm Selection 37](#_Toc79951879)

[3.2.2.2 Generating Correlation Matrix against Target Feature 39](#_Toc79951880)

[3.2.2.3 Iterative Training and Evaluation Technique 40](#_Toc79951881)

[3.2.2.4 Sorting and Displaying Model Performance 41](#_Toc79951882)

[Chapter 4: Result Analysis 43](#_Toc79951883)

[4.1 Model Rankings 43](#_Toc79951884)

[4.2 Top Model from Each Algorithm 45](#_Toc79951885)

[4.3 Round-up 48](#_Toc79951886)

[Chapter 5: Conclusion 50](#_Toc79951887)

[References 52](#_Toc79951888)

|  |
| --- |
| List of Tables |

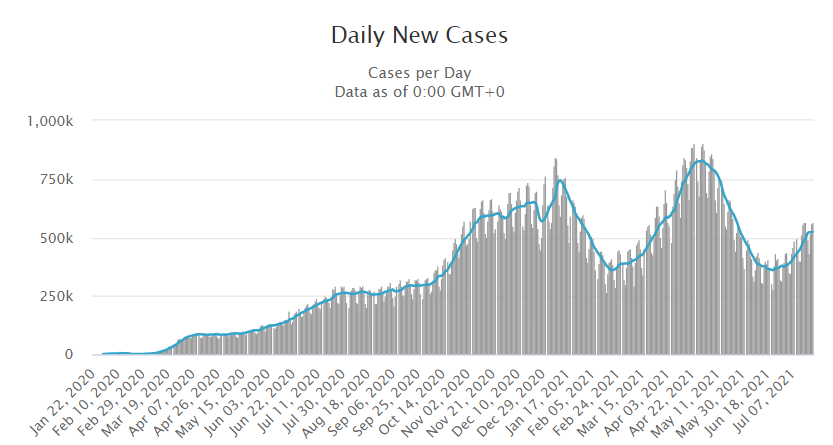
|  |  |  |
| --- | --- | --- |
| **Table 3-A** | Question & Response Charts | 21 |
| **Table 3-B** | Question-Answer to Feature-Value transformation | 32 |
| **Table 3-C** | Before Extraction | 33 |
| **Table 3-D** | After Extraction. | 33 |
| **Table 3-E** | Before One-Hot Encoding | 34 |
| **Table 3-F** | After One-Hot Encoding | 34 |
| **Table 4-A** | A brief ranking of the models | 43 |
| **Table 4-B** | Classification Report of Keras Model with 100 Features | 45 |
| **Table 4-C** | Classification Report for XGBoost Classifier Model with 60 Features | 46 |
| **Table 4-D** | Classification Report for Logistic Regression Model with 100 Features | 47 |
| **Table 4-E** | Full Performance Metrics of the best models from each algorithm | 48 |

|  |
| --- |
| List of Figures |

|  |  |  |
| --- | --- | --- |
| **Fig. 1-1** | Covid-19 Daily Cases since Detected [3] | 10 |
| **Fig. 1-2** | Covid-19 Symptoms Effectiveness [6] | 12 |
| **Fig. 3-1** | High-Level Steps Visualization of This Study | 31 |
| **Fig. 3-2** | A sample question from the survey form for describing One-Hot Encoding | 34 |
| **Fig. 3-3** | The formula of Min-Max Scaling | 35 |
| **Fig. 3-4** | Count of distinct observations in target feature before balancing | 35 |
| **Fig. 3-5** | Count of distinct observations in target feature after balancing | 36 |
| **Fig. 3-6** | Code Snippet Neural Network Keras Model | 38 |
| **Fig. 3-7** | High-level flow chart of iterative training and evaluation process | 42 |
| **Fig. 4-1** | Confusion Matrix for Keras Model with 100 Features | 45 |
| **Fig. 4-2** | Confusion Matrix for XGBoost Classifier Model with 60 Features | 46 |
| **Fig. 4-3** | Confusion Matrix for Logistic Regression Model with 100 Features | 47 |

|  |
| --- |
| Chapter 1: Introduction |

Covid-19 is a respiratory infection disease that can finally result in mortality. People have been suffering from this disease for the last one and a half years, and it has turned into a great pandemic of this century. A virus named Novel corona or SARS-CoV-2 is responsible for Covid-19. Coronavirus was first detected in Wuhan, China, in December 2019 [1]. This disease has infected more than 193 million people and killed above 4 million people worldwide [2]. The world has stopped for one and a half years. Educational institutions are closed for the last 13 months in different countries worldwide, creating a tremendous negative impact among the school & college students. Low-income people suffer from indigence as they are not getting enough work to maintain their daily living costs.



**Fig. 1-1**: Covid-19 Daily Cases since Detected [3]

Since its detection to the middle period, there were near about 250k daily cases worldwide on average. But after that period, the number of daily cases increased rapidly. The indifference of 3-4 months, the patients became triple of before. In Last May 2021, there was the highest number of daily cases, near about 1 million worldwide.

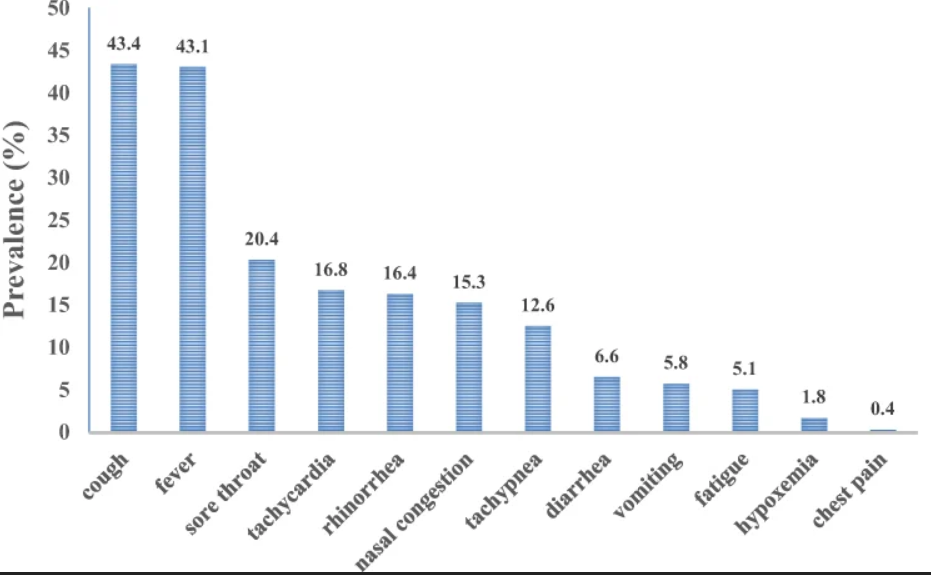
## The motivation of the study

The worse news is that the disease continuously changes its genotype and comes with different waves in different countries. At present, the virus is showing its brutality by continuing the second wave worldwide of its tarryingness. Specialists are predicting the virus will also come with a third wave. The world is uncertain because international trades are off, flights are off even though no one can move freely in their town. Statistics show this virus is more deadly than the previous century's Ebola virus, with the highest death rate during the period [4].

However, People and governments of their respective countries are trying their level best to eradicate this virus from everyday life. In this situation, the world will be benefited if we have a fast detection method for this disease. A whole community can be saved if they can predict this disease more accurately by analyzing previous data of infected patients of this disease. This research aims to predict the chances that a person will be affected by this disease by measuring their lifestyle data. These results will help future researchers to get more accurate data by data analysis.

## Symptom Analysis

Initially, the disease had identifiable symptoms like fever, dry cough, body pain, tiredness, etc. But with the increase of its spreading capability, the disease shows fewer symptoms of its effectiveness. Significantly, the Brazilian variant is more deadly with the highest death rate because this spread more quickly and cause a dire situation in a short time [5].



**Fig. 1-2**: Covid-19 Symptoms Effectiveness [6].

The most accurate symptoms were fever and cough. Tachycardia, Rhinorrhea, Nasal Congestion were moderate correct symptoms for Covid-19. But as time passes, the symptoms are getting blur. Covid-19 positive people have no symptoms where they are spreading the disease unlawfully. That's why people need fast analysis methods of suspicious symptoms.

.

## Lifestyle Analysis

Before recognizing this virus, people were free to live their lives according to their lifestyle. But as soon as the virus spreads worldwide, people are restricted from living an unclamped life. The daily habits that people have been following since the century are suddenly determined. Nowadays, people can not move without a mask, share food with their close ones, or even rub their own eyes. However, These restrictions are not harmful. For the benefit of humanity, several limitations have been imposed.

Whenever people are aware of their facemask, they are more protected from the coronavirus. Avoiding public gatherings keeps a human safe from any virus that exists. Sharing food with someone provides higher risks of spreading this virus. That's why people are developing new habits of avoiding all these risk factors.

All the governments of the country and the World Health Organization (WHO) are giving instructions and taking initiatives for creating a healthy lifestyle [7]. They are teaching people which habits should be avoided for a good life. Overpopulated countries are at significant risk as they sometimes fail to execute proper instructions as they could not maintain social distance. The governments are trying as hard much as they can.

Using online surveys, we can collect people's daily lifestyle data compared with the previous data. Covid-19 positive patient lifestyle data can be analyzed. The relevancy of lifestyle patterns with the disease can also be determined if researchers can easily differentiate the lifestyle facts those covid-19 positive patients generally had. Then those everyday lifestyle habits can be marked as avoidable for all people to stay away from coronavirus. By analyzing Covid-19 unhealthy lifestyle, we can easily detect the facts that are not responsible for spreading Covid-19.

|  |
| --- |
| Chapter 2: Literature Review |

A literature review summarizes, assesses, and distinguishes the current body of knowledge in a specific field of study. The source of information serves as a kind of blueprint, illustrating how research on the topic has progressed and outlining something that has already resulted, what is widely stated as fact; the current ability to comprehend the issue is still in its initial stages of development. This study area or case is referred to as literature, as are journals and published papers containing relevant conceptual documentation. It is not only a list of relevant sources but a review of the literature. When we evaluated sources and compiled the findings, we verified individual pieces of information and combined these elements to obtain a broader perspective on the topic. This chapter describes evolution, patterns, strategies, and points of contention and disagreement within this particular "field." To this end, we will explain our research and contribute to ongoing debates in the field. First and foremost, we read numerous papers, such as articles and journals, related to our chosen thesis topic. We primarily concentrated on specific parameters (Covid-19 symptom, covid-19 based case prediction).

Kolla Bhanu Prakash, S. Sagar Imambi, and other writers[8] represent an analysis, prediction, and evaluation of COVID-19 datasets using Machine Learning Algorithms. They have examined COVID-19 and made a dataset to determine which age groups are most affected by COVID-19. They used Machine learning techniques to create various prediction models such as Random Forest Regressor and Random Forest Classifier, SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression, and XGBoost Classifier, and they computed the performances.[8] The most affected age group are 20-50, who are suffered mainly by COVID -19. The correlation matrices are used to figure out how the features of the datasets are related. For the classifiers, the feature is calculated. Prediction is measured with the combination of classifiers and repressors. They took the dataset anonymously from Kaggle. The limitation is that they used only two models- Random Forest Regressor and Random Forest Classifiers to measure the accuracy but could not measure for Random Forest Classifiers.

Shuo Wang, Yunfei Zha, and their team [9] proposed a fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis by routinely using Computed Tomology. They used a two-step strategy to train the data.[9]They used a large dataset and computed CT images from 5372 patients gathered temporally from seven cities or provinces. First, they used 4106 patients' added tomography pictures to pre-train the deep learning system. The remaining patients were enrolled in a dataset with randomly chosen data for training and validation twice. After validation set 2, the accuracy rate was 85%, and they verified it externally with the help of the deep learning system. Their limitations are that they did not consider death cases or intensive care unit cases and did not look into severe and mild patients independently.

Kenji Ikemura, MD; Eran Bellin, and their team [10] proposed to use Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study. They presented Machine learning methods as a way to forecast the severity of COVID-19 disease. Their previous research has tested only one machine learning algorithm and evaluated efficiency using the areas under the curve method[10]. They chose the method which most accurately predicted the probability of patients surviving a covid outbreak. They also determined that specific parameters had the most significant impact on the accuracy of the model. To formulate the models, they used data from 4313 patients with 48 variables. Patients were examined for 30 days or until death. To acquire the most significant variable, they performed dimension reduction. The best independent models were the gradient boost machine, and extreme gradient boost models, and their accuracy rate is 0.803 and 0.793, respectively. After retraining the autoML models with these ten variables, they got the rate of 0.791. According to their goal, the concept of AutoML was found out as an exciting way to predict the chances of a patient's mortality.

Sonia Raf, Guillaume Bouzillé, and the other writers [11] proposed that Machine learning is the key to diagnose COVID 19:a proof of concept study[11]. The goal was to create and test machine learning models based on daily clinical and laboratory data to enhance RT-PCR's effectiveness and chest-CT for COVID-19 detection in patients hospitalized. They gathered data from 5196 patients to train the dataset but only used 536 suspected covid patients aged 18 or younger. They used three models to detect COVID-19: Binary Logistic Regression, Random Forest, and Artificial Neural Networks, each with 22 parameters. The AUCs for chest-CT and RT-PCR alone in diagnosing COVID-19 were 0.778 and 0.852, respectively. The limitation is that they used this method on only hospitalized patients.

Mohamed Abd Elaziz and his teams [12] present New machine learning method for image-based diagnosis of COVID-19. They developed a novel machine-learning algorithm to distinguish between COVID-19 patients and non-COVID-19 patients, assigning each set of chest x-ray images to one of two categories. Their segmentation technique incorporates two methodologies, presenting accurate COVID19 chest x-ray images [12]. They created two datasets called Dataset 1 and Dataset 2. Dataset 1 had 216 COVID-19 positive images of patients one to five years old, and Dataset 2 had 219 COVID-19 positive images and one 341 negative COVID-19 images of patients 40 to 84 years of age. They utilized the photos to identify features of datasets, and they subsequently devised a new method to optimize MRFO performance that focuses on employing differential equations. For both datasets, the accuracy of using the derived features without the feature selection approach is 0.901 and 0.9309, respectively.

Federico Cabitza, Andrea Campagner, and others[13] proposed that machine-learning models for detecting COVID-19, based on regular blood tests, should be developed, evaluated, and validated using three independent training datasets. They built three different data sets to train data with a model of machine learning. They collected 1624 patients data from various hospitals, and the majority were positive with Covid-19[13]. The dataset consists of up to 55 features. Five Machine Learning models(Logistic regression, Naïve Bayes, KNN, Random Forest, and SVM) for use in blood tests were developed successfully. Their accuracy rate for the three datasets varies from 0.75 to 0.78. They used this approach to identify positive Covid-19 patients who require faster and cheaper testing.

Matt J. Keeling, Edward M. Hill, and the other authors [14] analyzed Predictions of COVID-19 dynamics in the UK: Short-term forecasting and analysis of potential exit strategies. COVID-19 outbreaks in the UK are actively implementing research policymaking through Numerical methods. Using factual information on positive samples of hospitals and death rates, they illustrate an observable age-structural concept that utilizes that data to calculate new predictions on the onset of an outbreak in ten UK regions. This model is being used to compare three different exit strategy options[14]. They ran multiple experiments to estimate the potential impact of various strategies to alleviate social distancing overtime on inpatient and acute care admissions and deaths. They investigated the effects of decreasing enforcement, protecting elderly citizens against upcoming outbreaks, using regionally-based triggers to promote rigid social distances and undiagnosed disbursement.

Some other authors from IEEE [15] represent an Interpretable Machine Learning for COVID-19: An Empirical Study on Severity Prediction Task. Using a database of 92 patients who had confirmed SARS-CoV-2 laboratory tests, the researchers trained the data to identify biomarkers indicative of infection severity prediction. The authors demonstrate their findings using the understanding of four machine learning models, including decision trees, random forests, gradient boosted trees, and neural networks using permutation feature importance[15]. Their findings concur with the latest clinical research on COVID-19 and other studies using dedicated models, which indicate that the virus causes severe infection and increases the risk of death in humans. Following that, they validated the methods using two large open datasets containing many confirmed hospital patients. They use model interpretation rather than a high-accuracy black-box model to reveal the critical early diagnosis indicators rapidly. Gradient boosted trees, and neural networks surpass all other methods regarding consistency on the test set following training. They present both patients' test results to doctors without indicating which patient is experiencing the most severe symptoms. Doctors use the COVID-19 diagnosis to make the decisions, which are entirely identical to their medical models. When the two datasets are combined, it can determine that older individuals are more susceptible to the Covid virus than younger individuals.

Gregory L. Watson and his fellow authors proposed a [16] Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. They use an epidemiological compartmental model with a Bayesian time series model and a random forest algorithm to predict COVID-19 [16]. The model uses a random forest algorithm to determine which COVID-19 cases and population characteristics are associated with mortality. The two models are attached to a separate compartmental model, which projects active topics and recovered patients. Their model includes three key elements. Case and death models are used in the compartmental model as transition functions. A new COVID-19 case was predicted using a model that simulates the velocity of the cumulative log cases. A death model that included random forests was built, predicting deaths for each state and day in the United States. Because lagged cases and deaths are unknown at future dates and times, this model is dependent on another model to forecast the spread and progression of COVID-19 among the populations of various US states. They trained the model on the case, and death data collected and then used it to forecast the 21 days. The outcome was favorable; finally, the forecasts correctly estimated the cumulative confirmed cases and the daily death counts for New York, Colorado, and Westantha.

Dac Nhuong Le, Velmurugan Subbiah Parvathy, and their fellow researchers proposed a method [17] IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID 19 diagnosis and classification. Covid -19 is a severe virus and has led to a global catastrophe inadequately. A novel IoT-enabled Depth - wise separable convolutional neural network (DWS-CNN) with a Deep support vector machine (DSVM) is presented for COVID-19 classification and prediction. To achieve this, the DWS-CNN team has suggested a framework that includes data acquisition, Gaussian filtering (GF) as preprocessing, feature extraction, and classification processes. In the beginning, the patient data will be collected and sent to the cloud server using IoT devices. The GF technique also eliminates noise in the image. COVID-19 class labels are obtained using CXR images in the DWS-CNN model. The DWSCNN model uses two stages for classification: training and testing. In the first stage, named "GF-based preprocessing," features are extracted, and categories are applied. The DWS-CNN model's simulation processes are examined with the CXR image dataset. The experimental results guaranteed that the DWS-CNN model would obtain superior results in classification, resulting in 98.54% accuracy and 99.06% accuracy on binary and multiclass types, respectively[17].

We found that the most significant differences between our study or analysis of the paper and the existing studies were the following:

There have been many studies conducted on Contagious diseases like COVID-19 in the field of machine learning. Almost all the studies were conducted on clinical data regarding the physical factors of the patients. Most of the Study uses machine learning to predict mortality or chance of getting affected by analyzing X-Ray/C.T. images, blood sample reports, etc. It seems that there haven't been enough studies that used machine learning to explore people's social habits and lifestyles to predict the chance or likelihood of getting affected by such diseases even though many social habits/ hygiene rules are being enforced in people's daily life to prevent the disease from spreading / mass transmitting. This study focused on that criterion, where peoples' daily lifestyle and hygiene levels will be analyzed rather than clinical data regarding physical conditions to predict whether a person is in danger of getting affected by the disease for their lifestyle. Apart from the type or base of the study, if the technical approach is considered, this study also shows some uniqueness there. Like two of the previous studies have used deep learning models. But both of them were used to analyze image type of data [9][11]. This study uses a deep learning neural network model on a dataset consisting of categorical and non-ordinal binary data. And another significant fact about this study is, features were not trimmed before training and evaluating the models; in fact, the study selected the best model for predicting the disease alongside determining the best subset of the primary dataset by selecting features to be included in the subset using an iterative approach.

|  |
| --- |
| Chapter 3: Methodology |

The outbreak of the coronavirus showed the importance of a hygienic lifestyle for infection prevention. The purpose of this study and experiment was to develop a system that can analyze people's lifestyles and hygiene levels and predict the risk of getting infected by such contagious diseases.

## Dataset Description

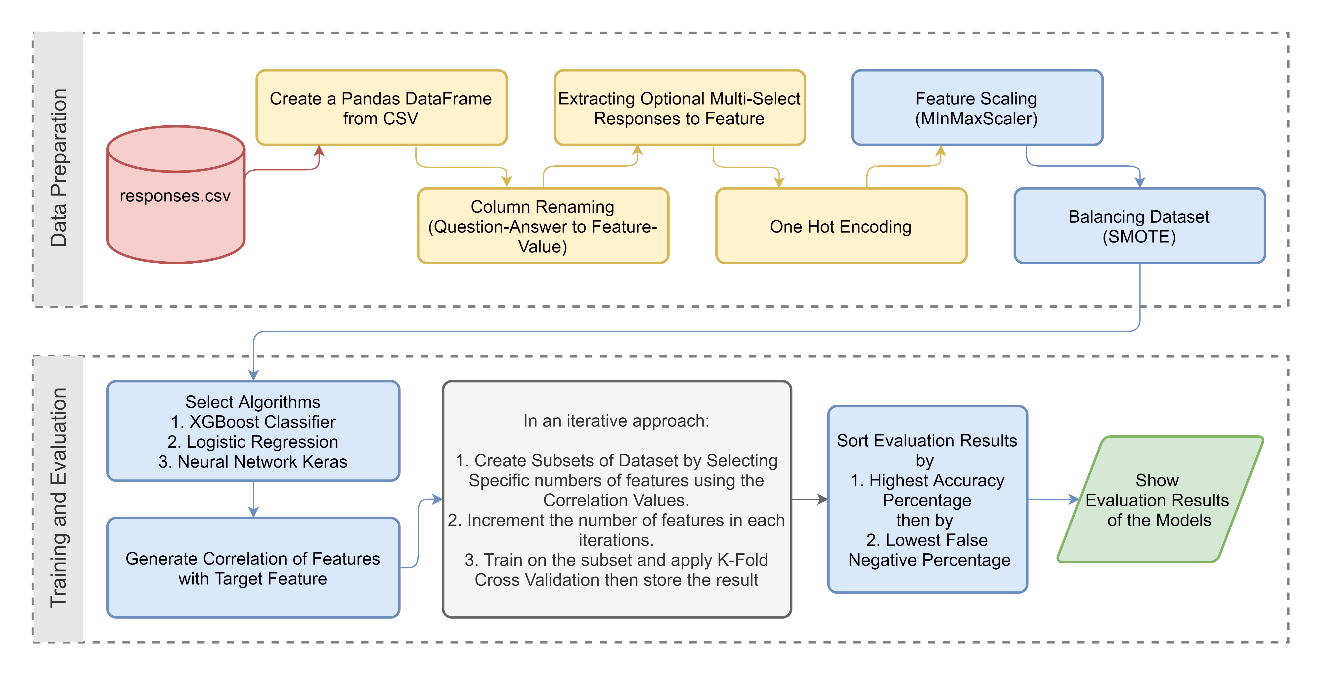
This study's approach required machine learning models to be trained with data depicting people's daily hygiene choices. Thus, a survey was conducted using a google form to collect data. A set of questions was prepared for the survey that reflect the lifestyle guidelines from the World Health Organization [18] and the Institute of Epidemiology Disease Control And Research [19], which is responsible for epidemiological and communicable disease research as well as the developing public health plans for the government of Bangladesh to implement. The question preparation required several meetings among all the members and the respected supervisor of this thesis. This phase eased the preprocessing and the feature engineering steps of this study, which will be discussed in later sections. The survey was done using a google form. In that form, there were 25 questions. Six hundred twenty responses were received before the start of the experiment with different algorithms. Table 3-A below visualizes all the questions and responses.

**Table 3-A:** Question & Response Charts

|  |  |
| --- | --- |
| SL | Question & Response |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |
| 11 |  |
| 12 |  |
| 13 |  |
| 14 |  |
| 15 |  |
| 16 |  |
| 17 |  |
| 18 |  |
| 19 |  |
| 20 |  |
| 21 |  |
| 22 |  |
| 23 |  |
| 24 |  |
| 25 |  |

## Machine Learning Approach

Several steps of the traditional machine learning approach were followed to develop this system. However, all the small steps can be categorized into two high-level abstract steps. The first one is Data Preparation, and the second one is Training and Evaluation. The high-level steps visualization can be seen from Fig. 3-1 below.



**Fig. 3-1:** High-Level Steps Visualization of This Study

### Data Preparation

#### Column Renaming

At the very beginning, right after creating a pandas dataframe from the responses, the column names were renamed into a 'feature-value form' from the 'question-answer form.' which can be seen from Table 3-B below.

**Table 3-B:** Question-Answer to Feature-Value transformation

|  |  |  |
| --- | --- | --- |
| **SL.** | **Question** | **Feature** |
| 1 | I am filling this form for – | FillingFor |
| 2 | Are you filling this form from Bangladesh? | FillingFromBD |
| 3 | Have you ever been diagnosed With COVID Positive? | CovidPositive |
| 4 | Have any of your family members / House-mates been diagnosed With COVID Positive? | RelatedCovidPositive |
| 5 | Have you taken any VACCINE for COVID-19? | Vaccinated |
| 6 | What is your living area type? | LivingAreaType |
| 7 | Your Profession | Profession |
| 8 | Your Job type | ProfessionType |
| 9 | Your Age (Years) [ example: 25 ] | Age |
| 10 | Your Body Weight (KG) [ example: 65 ] | Weight |
| 11 | Your gender | Gender |
| 12 | Tell us about your accommodation type | AccommodationType |
| 13 | How many people do live with you in your house? | HouseMateCount |
| 14 | Are you diagnosed with any of the diseases below? | Disease |
| 15 | Do you wear face masks? | WearingFaceMask |
| 16 | Do you uncover your mouth while talking to others? | UncoveringMouth |
| 17 | How frequently do you have to go outside? | GoingOutside |
| 18 | How frequently do your family members / House-mates go outside? | RelatedGoingOutside |
| 19 | For what reason do you go outside often? | GoingOutsideReason |
| 20 | Do you use hand sanitizer? | UsingSanitizer |
| 21 | How frequently do you wash your hands? | WashingHands |
| 22 | How frequently do you have to dine outside during pandemic? (restaurants, hotel, workplace) | DiningOutside |
| 23 | Do you maintain social distance when you go outside? | MaintainingSocialDistance |
| 24 | How frequently do you use public transport? | UsingPublicTransport |
| 25 | How frequently do your friends or family members come to your house for visiting purposes? | VisitByFnF |

#### Extraction of Columns

Checkboxes were used to receive multiple answers for two of the questions. Those multiple answers were stored in cells in a semicolon-separated manner. Python's 'str.get\_dummies' method was used to extract those values. Then they were appended as binary columns in the dataframe. From below two tables, it can be understood where Table3-C represents the before extraction scenario, and Table 3-D shows the after extraction scenario.

**Table 3-C**: Before Extraction

|  |  |
| --- | --- |
|  | Disease |
| n | High Blood-Pressure; Allergy |

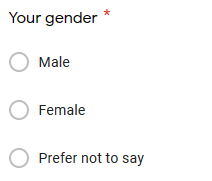
**Table 3-D**: After Extraction

|  |  |  |
| --- | --- | --- |
|  | Disease\_High Blood-Pressure | Disease\_Allergy |
| n | 1 | 1 |

#### One-Hot Encoding

Some algorithms can work with categorical data directly. Like decision trees can be trained with categorical data directly. But many machine learning algorithms require a numerical representation of data to train models. For this reason, the response values (especially the responses submitted through radio buttons) were transformed from categorical to numerical representation. The transformation process was One-Hot Encoding in this case. The reason for using One-Hot Encoding is that the values from those features are not ordinal. If they were ordinal, level encoding, or, to be specific, the 'integer encoding' technique could also be used. One-Hot encoding converts categorical information into a format that may be fed into machine learning algorithms to improve prediction accuracy.

For Example:



**Fig. 3-2:** A sample question from the survey form for describing One-Hot Encoding

If anybody responded with 'Male' for the question visible in Fig. 3-2, generated a row like Table 3-E, and preprocessing the data would convert it like Table 3-F.

**Table 3-E**: Before One-Hot Encoding

|  |  |
| --- | --- |
|  | Your gender |
| n | Male |

**Table 3-F**: After One-Hot Encoding

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gender\_Male | Gender\_Female | Gender\_Prefer not to say |
| n | 1 | 0 | 0 |

#### Scaling Features

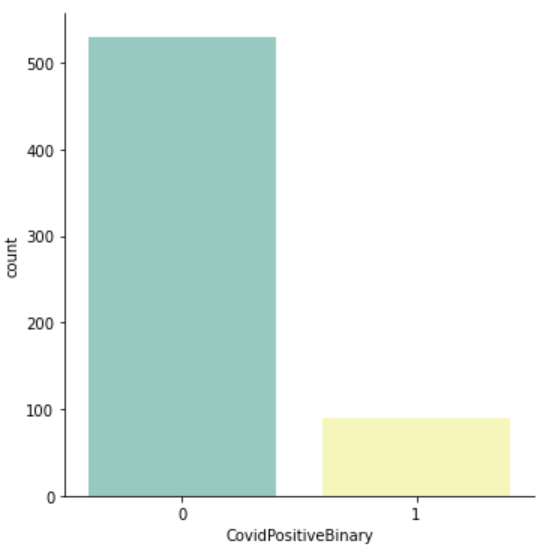
Suppose a feature's dispersion is orders of magnitude more than the dispersion of other features. In that case, that feature may overpower other features from the dataset, which is not wanted in this study's modeling. Thus the feature values of the dataset were scaled to keep them in a specific range. Min-Max Scaling was used in this approach. This scaling will ensure that the feature values be in the range of 0 to 1. The formula for this approach can be seen in Fig. 3-3.



**Fig. 3-3:** The formula of Min-Max Scaling.

#### Balancing Dataset

There are situations in machine learning where the number of observations in one class is significantly lower than the number of observations in the other classes. The dataset created from the survey responses for this study had only 90 responses with covid positive diagnosis affirmative out of 620 responses. The visualization of this can be seen in Fig. 3-4 below.

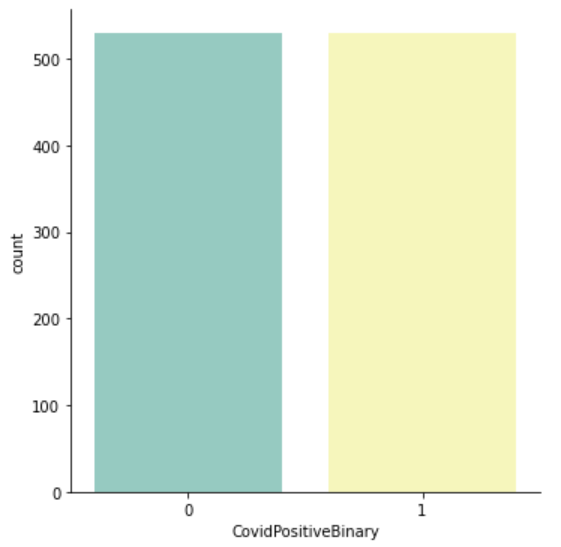


**Fig. 3-4:** Count of distinct observations in target feature before balancing.

Thus the dataset needed to be balanced to have an almost equal amount of positive (1) and negative (0) responses for the 'CovidPositiveBinary' column.

So, Synthetic Minority Oversampling Technique (SMOTE) was used for balancing the dataset. SMOTE works by selecting instances in the feature space that are close together, drawing a line in the feature space between the examples, and drawing a new sample at a location along that line [20].

After balancing the dataset, there were equal responses for covid positive and negative cases, and it can be seen in the visualization from Fig. 3-5.



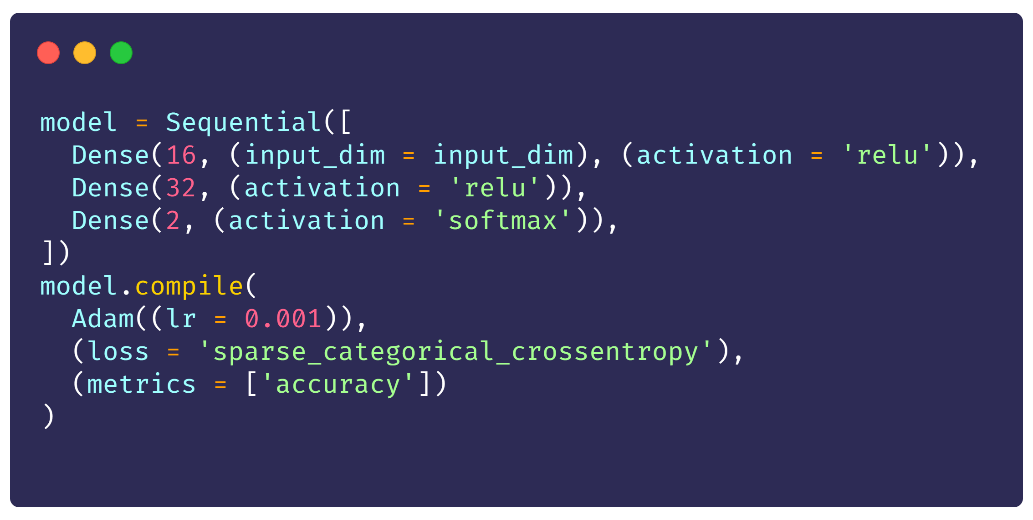
**Fig. 3-5:** Count of distinct observations in target feature after balancing.

### Training and Evaluation

#### Algorithm Selection

Three popular machine learning algorithms were used to train models for this approach:

1. **XGBoost Classifier:** The name XGBoost stands for Extreme Gradient Boosting. Speed and efficiency are the goals of XGBoost, an implementation of gradient-boosted decision trees. XGBoost is a machine learning algorithm that has lately dominated the field of applied machine learning, according to current research. The XGBoost algorithm was created at the University of Washington as part of a research effort. During the 2016 SIGKDD Conference, Tianqi Chen and Carlos Guestrin presented a work that ignited the Machine Learning community [21]. It has won numerous Kaggle competitions and is the brains behind several cutting-edge industry applications. XGBoost is a fast algorithm. When it is compared to other implementations of gradient boosting, this one is very fast. When it comes to classification and regression tasks, XGBoost dominates tabular and structured datasets.
2. **Logistic Regression:** When there are one or more independent variables,  Logistic Regression is a fairly basic classifier to predict the output. In binary form, the output value maybe 0 or 1. Because it's simpler to implement, understand, extremely fast to train, and readily extendable to many classes (multinomial regression), Logistic Regression was chosen. It also offers a natural probabilistic perspective of class predictions, which one may find appealing. The Logistic Regression method requires less training time than other complicated algorithms because of its straightforward probabilistic interpretation. Other more complicated Algorithms may also be measured using Logistic Regression as a baseline. If the dataset is low-dimensional, Logistic Regression can also handle it. In addition, it is less prone to over-fitting and does not need a large number of computing resources.
3. **Neural Network Keras:** As far, two basic algorithms were chosen from a statistical domain. But this one was considered to push the limit a bit more. For neural network modeling, Keras was the pick [22]. Free and open-source Python package Keras is a powerful and easy-to-use tool for constructing and analyzing deep learning models. As a result, one can construct and train neural network models in a few lines of code. It wraps libraries like Theano and TensorFlow, which are known for efficient numerical computation.



**Fig. 3-6:** Code Snippet Neural Network Keras Model

As seen from Fig. 3-6, a sequential approach, stacking three tensor layers, was used for this approach. The activation function ReLU was used for the first two layers**,** and the third or output layer, the **Softmax** function, was used.

ReLU or the rectified linear unit is an activation function that outputs the input straight if it is positive; otherwise, it outputs zero. For many types of neural networks, it has become the default activation function because it is a model that is quicker to train and typically performs better [23].

When converting a vector of integers to an equally-spaced vector of probabilities, the probability of each value is proportional to the vector's relative scale. Softmax is a mathematical function that performs this conversion. The function is most commonly used in applied machine learning as an activation function in a neural network model, which is the most popular use of the function. Furthermore, the network is designed to output N values, one for each of the classes in the classification task, and the softmax function is used to normalize the outputs, turning them from weighted sum values to probabilities that add up to 1. Specifically, each number returned by the softmax function can be understood as the chance of belonging to each class in the function's output [24].

For the compilation, **Adam** with the **sparse\_categorical\_crossentropy** was used as the loss function. In the optimization technique, Adam optimization is a stochastic gradient descent method based on the adaptive estimate of first-order and second-order moments [25].

#### Generating Correlation Matrix against Target Feature

It was necessary to see how many features were needed to gain a specific level of accuracy to find out the most suitable algorithm and model combination for this study and similar machine learning jobs. Selection of features in such a manner was only possible if a ranking could be established for all the features against the target feature 'CovidPositiveBinary.' This feature importance ranking can be achieved using several mechanisms. For this Study, Correlation Matrix was used to analyze feature importance and rank them accordingly.

#### Iterative Training and Evaluation Technique

After generating the correlation matrix, a dictionary (python data structure) was created for storing the evaluation results. Then a variable named 'subset' was declared with an initial value of 60. Then inside a loop, a subset of the dataset was created where feature count was equal to the value of the subset variable (for the first iteration, the sub-dataset contained 60 most important features according to the correlation matrix for the target feature. With that target feature included, the sub-dataset had 61 columns). A model was then created for each of the three selected algorithms right after that. Then those models were trained on the sub-dataset. After training each model with that supplied sub-dataset, K-Fold cross-validation was applied on each model, where the value of K was 10.

Cross-validation is a resampling technique for evaluating machine learning models on a small sample of data. The process includes only one parameter, k, which specifies the number of groups into which a given data sample should be divided. So it's called k-fold cross-validation. When a precise value for k is specified, it can be substituted for k in the model's reference, for example, k=10 for 10-fold cross-validation. Cross-validation is a technique used in applied machine learning to estimate a machine learning model's skill on unknown data. Using a small sample to assess how the model will perform in general when used to generate predictions on data that was not utilized during the model's training. It's a popular strategy since it's straightforward to grasp and produces a less biased or optimistic estimate of model competence than other approaches, such as a simple train/test split.

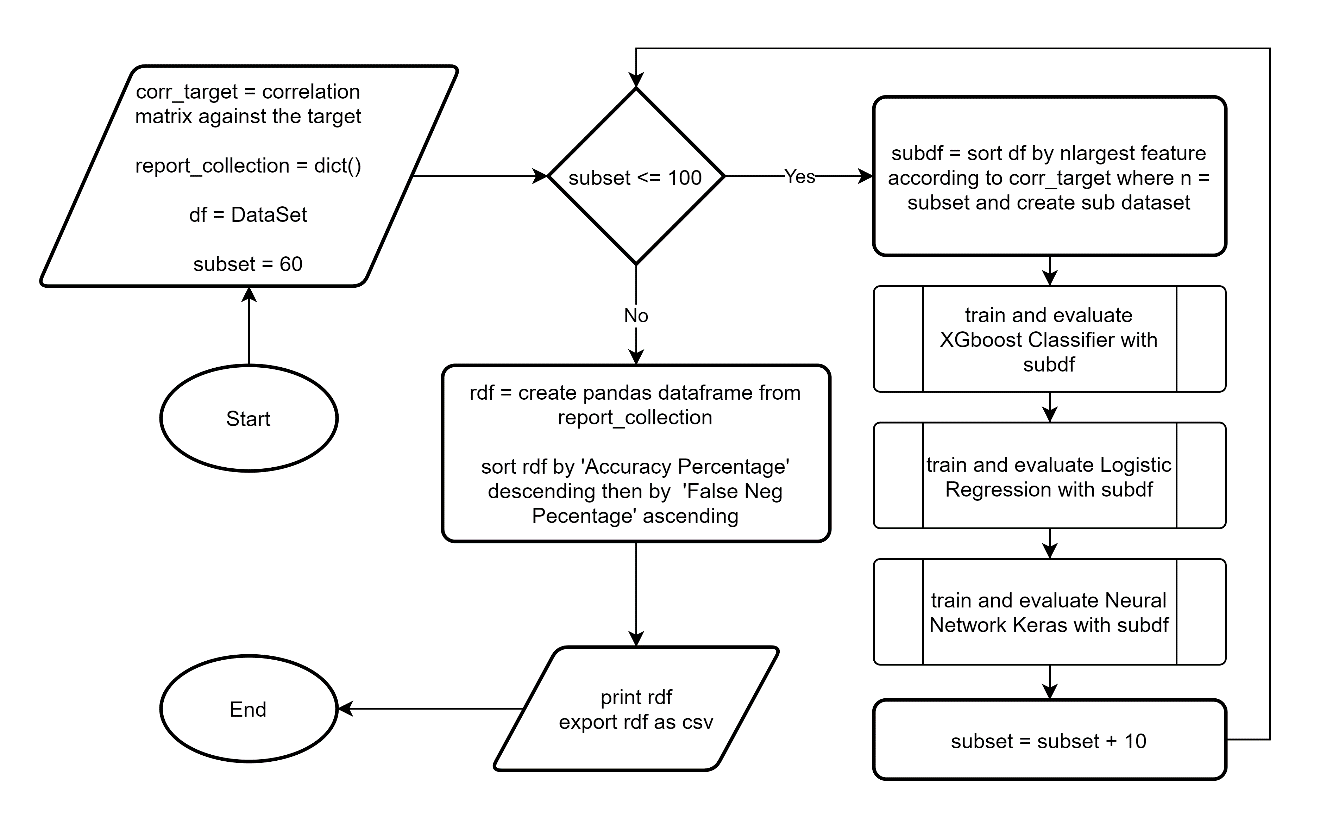
The value of 'subset,' which was used to specify the number of features included in the sub-dataset, was incremented by 10 at the end of each iteration. The full dataset had 100 features (excluding the target feature), the loop ran five times before breaking out. Thus, the dictionary used to store evaluation results had a performance report for 15 models (3 algorithms trained with five different sub-datasets) by the end of 5 iterations. The performance metrics: Accuracy, Precision, F1 Score, and Recall were generated from cross-validation results for each model using the 'classification\_report' function. The function is from the 'metrics' module from the 'sklearn' library for python [26]. Those values, alongside the confusion matrix values, were stored in the dictionary for every model.

Because of the stochastic nature of the neural network algorithm and K-Fold evaluation procedure for the algorithm, the results can vary or show differences in numeric precision in each run [27]. It could affect the ranking as well. Hence, to stabilize the order to some extent more, the K-Fold cross-validation was applied 10 times on each model trained with Keras, and the average was considered and stored as the final result. Due to this step, some metrics value like true/false positive/negative count was stored in the dictionary as floating-point numbers instead of plain round integers.

#### Sorting and Displaying Model Performance

After completing all the iteration, a pandas dataframe from the result storage dictionary was created. Then the dataframe was sorted by the models' 'Accuracy Percentage' and then by the 'False Neg Percentage.' After printing the dataframe, it was also exported as a CSV file for future usage.

The below Fig. 3-7 shows a flow chart that describes the whole iterative flow of the training and evaluating phase.



**Fig. 3-7:** High-level flow chart of iterative training and evaluation process

|  |
| --- |
| Chapter 4: Result Analysis |

## Model Rankings

As previously mentioned, all the three selected algorithms were trained for 5 sets of datasets containing feature counts from 60 to 100. It resulted in a performance report collection for 15 models (5 models for each algorithm). Despite the stabilization step (described at the end of section [3.2.2.3](#_Iterative_Training_and)), the ranking for neural network models varied for every run of Jupyter Notebook due to the slight change of accuracy percentage. The result-set obtained from the last run of the notebook is discussed in this paper.

**Table 4-A:** A brief ranking of the models

|  |  |  |
| --- | --- | --- |
| **Rank** | **Algorithm** | **Accuracy Percentage** |
| 1 | Keras 100 | 94.24 |
| 2 | Keras 90 | 94.2 |
| 3 | Keras 60 | 94.14 |
| 4 | Keras 70 | 94.12 |
| 5 | Keras 80 | 94.03 |
| 6 | XGBoost Classifier 60 | 92.64 |
| 7 | XGBoost Classifier 90 | 92.64 |
| 8 | XGBoost Classifier 70 | 92.55 |
| 9 | XGBoost Classifier 80 | 92.08 |
| 10 | XGBoost Classifier 100 | 91.79 |
| 11 | Logistic Regression 100 | 85.57 |
| 12 | Logistic Regression 70 | 84.81 |
| 13 | Logistic Regression 60 | 84.72 |
| 14 | Logistic Regression 80 | 84.62 |
| 15 | Logistic Regression 90 | 84.25 |

Table 4-A clarifies one thing, models with the Neural Network Keras algorithm outperformed the models with other algorithms. Even the lowest performer model with neural network Keras has an accuracy of 94.03%, which is higher by 1.93% percent than the best non-neural network algorithm.

## Top Model from Each Algorithm

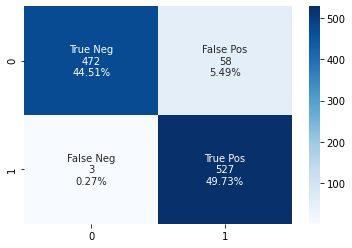
**Keras 100:**

***Algorithm:*** Neural Network Keras

***Feature Count (excluding target):*** 100

**Table 4-B:** Classification Report ofKeras Model with 100 Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1 Score** | **Accuracy** |
| 0 | 0.993891374 | 0.890188679 | 0.939163055 | 0.942358491 |
| 1 | 0.900621613 | 0.994528302 | 0.945232999 |
| Average | 0.947256493 | 0.942358491 | 0.942198027 |



**Fig. 4-1:** Confusion Matrix for Keras Model with 100 Features

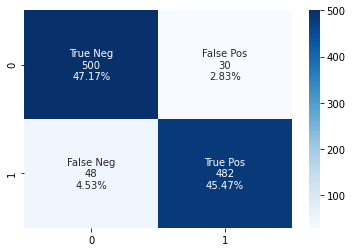
**XGboost Classifier 60:**

***Algorithm:*** XGBoost Classifier

***Feature Count (excluding target):*** 60

**Table 4-C:** Classification Report for XGBoost Classifier Model with 60 Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1 Score** | **Accuracy** |
| 0 | 0.912408759 | 0.943396226 | 0.927643785 | 0.926415094 |
| 1 | 0.94140625 | 0.909433962 | 0.925143954 |
| Average | 0.926907505 | 0.926415094 | 0.926393869 |



**Fig. 4-2:** Confusion Matrix for Neural Network Keras Model with 100 Features

As seen from the report, this one is the best model if neural network models are excluded from the ranking. It outperformed the other non-neural algorithm, 'Logistic Regression,' by a considerable margin. One noteworthy point about this model is that XGBoost Classifier performs best when feeding it 60 features only, which means it is doing its best when provided with 40% fewer features. This model's performance is impressive, keeping in mind that it is way less resource-hungry than any neural network model.

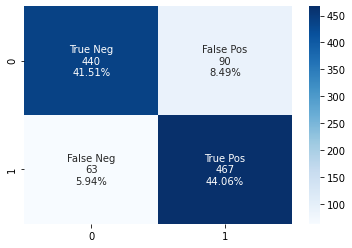
**Logistic Regression (100 Feature Count):**

***Algorithm:*** Logistic Regression

***Feature Count (excluding target):*** 100

**Table 4-D:** Classification Report for Logistic Regression Model with 100 Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1 Score** | **Accuracy** |
| 0 | 0.874751491 | 0.830188679 | 0.851887706 | 0.855660377 |
| 1 | 0.838420108 | 0.881132075 | 0.85924563 |
| Average | 0.856585799 | 0.855660377 | 0.855566668 |



**Fig. 4-3:** Confusion Matrix for Logistic Regression Model with 100 Features

As it can be seen, models with the Logistic Regression algorithm as a classifier came out to be the worst performers in this study. This one is the best model trained with Logistic Regression with 85.57% accuracy, 7.07% less than the best model trained with another non-neural algorithm XGBoost Classifier.

## Round-up

**Table 4-E:** Full Performance Metrics of the best models from each algorithm.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Keras softmax 100** | **XGBoost Classifier 60** | **Logistic Regression 100** |
| Accuracy | 0.942358491 | 0.926415094 | 0.855660377 |
| True Neg Count | 471.8 | 500 | 440 |
| True Neg Percentage | 0.44509434 | 0.471698113 | 0.41509434 |
| False Pos Count | 58.2 | 30 | 90 |
| False Pos Percentage | 0.05490566 | 0.028301887 | 0.08490566 |
| False Neg Count | 2.9 | 48 | 63 |
| False Neg Percentage | 0.002735849 | 0.045283019 | 0.059433962 |
| True Pos Count | 527.1 | 482 | 467 |
| True Pos Percentage | 0.497264151 | 0.454716981 | 0.440566038 |
| Precision\_0 | 0.993891374 | 0.912408759 | 0.874751491 |
| Precision\_1 | 0.900621613 | 0.94140625 | 0.838420108 |
| Precision\_macro\_avg | 0.947256493 | 0.926907505 | 0.856585799 |
| Precision\_weighted\_avg | 0.947256493 | 0.926907505 | 0.856585799 |
| Recall\_0 | 0.890188679 | 0.943396226 | 0.830188679 |
| Recall\_1 | 0.994528302 | 0.909433962 | 0.881132075 |
| Recall\_macro\_avg | 0.942358491 | 0.926415094 | 0.855660377 |
| Recall\_weighted\_avg | 0.942358491 | 0.926415094 | 0.855660377 |
| F1-Score\_0 | 0.939163055 | 0.927643785 | 0.851887706 |
| F1-Score\_1 | 0.945232999 | 0.925143954 | 0.85924563 |
| F1-Score\_macro\_avg | 0.942198027 | 0.926393869 | 0.855566668 |
| F1-Score\_weighted\_avg | 0.942198027 | 0.926393869 | 0.855566668 |
| Support\_0 | 530 | 530 | 530 |
| Support\_1 | 530 | 530 | 530 |
| Support\_macro\_avg | 1060 | 1060 | 1060 |
| Support\_weighted\_avg | 1060 | 1060 | 1060 |

Here, it can be seen from the table cell highlights that Keras softmax with 100 Feature count is winning almost every metric of the performance. The metric F1 score is backing it up firmly, even though it has lost to XGBoost Classifier in terms of true-negative detection and false-positive detection criteria.

According to the performance of the top model, the main goal of this study seems accomplished with an acceptable accuracy level, which was to develop a machine learning system that will analyze people's daily lifestyle and hygiene levels and predict if they are affected or going to be affected by contagious diseases like COVID-19.

The relevancy of a hygienic lifestyle is also validated as all the models of this study showing consistency and a decent level of accuracy. The prediction based on people's lifestyle wouldn't show this much accuracy if lifestyle did not matter.

|  |
| --- |
| Chapter 5: Conclusion |

The outbreak of coronavirus, a contagious disease, has shown its terrible effects on people's daily life. However, studies on the disease have also shown that diseases like this can be prevented from mass transmission if people maintain a regulated and recommended hygienic lifestyle. Hence, it was a necessity to analyze and observe the effectiveness of such claims. This study has done such analysis from a machine learning perspective under the umbrella of computer science. A public survey regarding lifestyle-related queries aligned with the guidelines from international organizations like WHO and national institutes like IEDCR has been conducted. The survey resulted in a dataset consisting of responses from 620 people. After that, in the study, traditional machine learning approaches have been taken. Data preprocessing like converting categorical data to numeric representation, feature engineering like scaling the data and stretching the dataset for balancing target class, etc., have been followed.

The study has incorporated an iterative approach instead of trimming the dataset based on feature importance before the training and evaluation phase. In that approach, three popular machine learning algorithms have been trained and evaluated with a range of sub-datasets that were prepared by selecting specific number features based on the feature importance according to correlation-matrix. The study has considered 15 models with two popular non-neural network algorithms and one neural network algorithm: XGBoost Classifier, Logistic Regression, and Neural Network Keras. It has been seen from the study that the deep learning model "Neural Network Keras" with a 100 feature count is the most successful model for this type of scenario. Gradient boosting classifier like XGBoost Classifier has also shown its capability with 40% fewer features in the experiment.

In conclusion, the study's goal has been accomplished. The selected machine learning model can analyze a person's lifestyle and predict if they will be affected by the disease with an acceptable accuracy rate of 94.24%. The relevancy of a hygienic lifestyle is also validated as the machine learning models that this study has prepared and evaluated have shown good accuracy. Yet, some limitations of this study can not be overlooked, like the study has been conducted on a dataset that was prepared upon a survey that consisted of voluntary responses from the people. These things question the integrity of the approach. If the study could have been operated in a controlled environment with clinically attested physical conditions of the subjects (people), it would be much more robust. In the future, with proper funding, such approaches can be taken. And one more thing is that ranking out habits/hygiene practices that enhance the risk of getting affected by such viruses could not be determined as it was getting out of the scope of this study. This can also be considered for future improvement.

|  |
| --- |
| References |

1. "About the virus", *Euro.who.int*. [Online]. Available: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov. [Accessed: 17- Jul- 2021].
2. "COVID Live Update: 205,137,769 Cases and 4,332,555 Deaths from the Coronavirus - Worldometer", *Worldometers.info*, 2021. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 17- Jul- 2021].
3. "COVID Live Update: 205,137,769 Cases and 4,332,555 Deaths from the Coronavirus - Worldometer", *Worldometers.info*, 2021. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 17- Jul- 2021].
4. M. Kortepeter, "Why Is Covid-19 More Deadly Than Ebola? An Infectious Disease Doctor Explains", *Forbes*, 2021. [Online]. Available: https://www.forbes.com/sites/coronavirusfrontlines/2020/07/31/why-is-covid-19-more-deadly-than-ebola-an-infectious-disease-doctor-explains/?sh=7f29edb9f734. [Accessed: 18- Jul- 2021].
5. "COVID-19 variant found in Brazil ‘spreads faster’ - SciDev.Net", *SciDev.Net*, 2021. [Online]. Available: https://www.scidev.net/global/news/covid-19-variant-found-in-brazil-spreads-faster/. [Accessed: 18- Jul- 2021].
6. H. Zare-Zardini, H. Soltaninejad, F. Ferdosian, A. Hamidieh and M. Memarpoor-Yazdi, "Coronavirus Disease 2019 (COVID-19) in Children: Prevalence, Diagnosis, Clinical Symptoms, and Treatment", *International Journal of General Medicine*, vol. 13, pp. 477-482, 2020. Available: 10.2147/ijgm.s262098 [Accessed 18- Jul- 2021].
7. "COVID-19 Mythbusters – World Health Organization", *Who.int*, 2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters?gclid=CjwKCAjwx8iIBhBwEiwA2quaqxGrl\_sRamCd902KiMRq-JpyB78GM0yj1BJtfwEzb62Lf\_80Nnqu1xoC7lkQAvD\_BwE#virus. [Accessed 18- Jul- 2021].
8. K. Prakash, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms", *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 2199-2204, 2020. Available: 10.30534/ijeter/2020/117852020 [Accessed: 19- Jul- 2021].
9. S. Wang et al., "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis", *European Respiratory Journal*, vol. 56, no. 2, p. 2000775, 2020. Available: 10.1183/13993003.00775-2020 [Accessed: 19- Jul- 2021].
10. K. Ikemura et al., "Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study", *Journal of Medical Internet Research*, vol. 23, no. 2, p. e23458, 2021. Available: 10.2196/23458 [Accessed: 19- Jul- 2021].
11. C. Gangloff, S. Rafi, G. Bouzillé, L. Soulat and M. Cuggia, "Machine learning is the key to diagnose COVID-19: a proof-of-concept study", *Scientific Reports*, vol. 11, no. 1, 2021. Available: 10.1038/s41598-021-86735-9 [Accessed: 20- Jul- 2021].
12. M. Elaziz, K. Hosny, A. Salah, M. Darwish, S. Lu and A. Sahlol, "New machine learning method for image-based diagnosis of COVID-19", *PLOS ONE*, vol. 15, no. 6, p. e0235187, 2020. Available: 10.1371/journal.pone.0235187 [Accessed: 20- Jul- 2021].
13. F. Cabitza et al., "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests", *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 59, no. 2, pp. 421-431, 2020. Available: 10.1515/cclm-2020-1294 [Accessed: 21- Jul- 2021].
14. M. Keeling et al., "Predictions of COVID-19 dynamics in the UK: Short-term forecasting and analysis of potential exit strategies", *PLOS Computational Biology*, vol. 17, no. 1, p. e1008619, 2021. Available: 10.1371/journal.pcbi.1008619 [Accessed: 21- Jul- 2021].
15. H. Wu et al., "Interpretable Machine Learning for COVID-19: An Empirical Study on Severity Prediction Task", *IEEE Transactions on Artificial Intelligence*, pp. 1-1, 2021. Available: 10.1109/tai.2021.3092698 [Accessed: 22- Jul- 2021].
16. G. Watson et al., "Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model", *PLOS Computational Biology*, vol. 17, no. 3, p. e1008837, 2021. Available: 10.1371/journal.pcbi.1008837 [Accessed: 22- Jul- 2021].
17. D. Le, V. Parvathy, D. Gupta, A. Khanna, J. Rodrigues and K. Shankar, "IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID-19 diagnosis and classification", *International Journal of Machine Learning and Cybernetics*, 2021. Available: 10.1007/s13042-020-01248-7 [Accessed: 23- Jul- 2021].
18. "Advice for the public on COVID-19 – World Health Organization", *Who.int*. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public. [Accessed: 26- Jul- 2021].
19. "COVID-19 General Information | IEDCR", *IEDCR*. [Online]. Available: https://iedcr.gov.bd/covid-19/covid-19-general-information. [Accessed: 26- Jul- 2021].
20. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. Available: 10.1613/jair.953.
21. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. Available: 10.1145/2939672.2939785 [Accessed 26 Jul 2021].
22. K. Team, "Keras documentation: Why choose Keras?", *Keras.io*, 2021. [Online]. Available: https://keras.io/why\_keras/. [Accessed: 27- Jul- 2021].
23. J. Brownlee, "A Gentle Introduction to the Rectified Linear Unit (ReLU)", *Machine Learning Mastery*, 2019. [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. [Accessed: 27- Jul- 2021].
24. J. Brownlee, "Softmax Activation Function with Python", *Machine Learning Mastery*, 2020. [Online]. Available: https://machinelearningmastery.com/softmax-activation-function-with-python/. [Accessed: 27- Jul- 2021].
25. K. Team, "Keras documentation: Adam", *Keras.io*. [Online]. Available: https://keras.io/api/optimizers/adam/. [Accessed: 27- Jul- 2021].
26. "sklearn.metrics.classification\_report — scikit-learn 0.24.2 documentation", *Scikit-learn.org*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification\_report.html. [Accessed: 28- Jul- 2021].
27. J. Brownlee, "Why Do I Get Different Results Each Time in Machine Learning?", *Machine Learning Mastery*. [Online]. Available: https://machinelearningmastery.com/different-results-each-time-in-machine-learning/. [Accessed: 28- Jul- 2021].