Journal of Applied Sciences, Information and Computing
Vol. 2, No. 1 (June, 2021)
School of Mathematics and Computing, Kampala International University

ISSN: 1813-3509 https://jasic.kiu.ac.ug

Development of Workforce Diseases Detection System Using Machine Learning Models

Emuoyibofarhe O.N [1], Famutim R. F [2]. Olanloye D.O.[3], Adebayo .S.[4], Abdulraheem A. Ali[5]
ozichi.emuoyibofarhe@bowen.edu.ng, ranti.famutimi@bowen.edu.ng, odun.olanloye@bowen.edu.ng, aliakolade@gmail.com, segun.adebayo@bowen.edu.ng

[1,2,3,5]College of Computing and Communication Studies, Bowen University, Iwo, Osun State, Nigeria,
[4]College of Agriculture, Engineering and Sciences, Bowen University, Iwo, Osun State, Nigeria,
,

Abstract

Health Status is one of the important parameters determining workers' productivity and, therefore, significantly affects the nation's workforce. This research work develops a workforce disease predictive system using Machine Learning Models. The models were trained using the five different machine learning algorithms, which include Neural Networks, Random Forest Classifier, XGBoost, Multilayer Perceptron, K-Nearest Neighbour, and uses insertion sort algorithm for the second tier, which sorts the results to analyse and select the most likely disease for each of the workers. Hence, a system is developed which is capable of detecting such workforce diseases. Python programming language with Scikit Learn and TensorFlow Library was used as an instrument for implementation. The system is more accurate and unique in predicting workforce diseases than other existing methods or techniques

Keywords: Workforce, Diseases, Machine Learning Models, Python Programming Language, Prediction.

1. Introduction

As the name implies, workforce refers to the number of persons who are working or, better still, the number of potentially assignable workers for any purpose in a global economy. According to World Health Organisation, in most developed countries each year, 12.2 million people die from non-communicable diseases at their active working age. The work-related health problems/challenges in most countries result in an economic loss of 4 to 6% of Gross Domestic Product. Workplace health hazards such as heat, noise, dust, toxic substances, unsafe equipment, and psychological stress may cause occupational disease.

Machine learning (ML) is the scientific study of mathematical models used by computing systems to perform specific functions without using specific instructions, focusing instead on patterns and inferences about the mathematical representation of the real world. Research has shown that healthy employees are productive and are expected to raise healthy families when satisfied fit in their day-to-day activities. In fact, a larger percentage of their illness is as a result of workforce-related diseases. Also, by creating a positive, safe, and healthy environment for employees, one can increase the employees' morale and, in turn, positively impact productivity at various levels. Some health risk factors can lead to workforce diseases which can aggravate other related health issues among the workers. Studies have shown that employees working under stress or with precarious employment conditions are likely to smoke more, exercise less, and have an unhealthy diet. Such calibre of people may be vulnerable to chronic respiratory diseases, musculoskeletal disorders or injuries, noise-induced hearing loss, carcinogenic agents, visual disorder, airborne particles, ergonomic risks, and Skin problems, which are the most common workforce diseases or occupational disease. Hence, a workforce disease is any chronic ailment that occurs because of work or occupational activities.

Machine learning (ML) is the scientific study of algorithms and mathematical models used in computing systems to perform specific functions without using specific instructions but focusing on patterns and inferences. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. To generate a machine learning model, functions that will predict output for a particular input will have to be formulated.

Health care organisations are leveraging machine-learning techniques such as artificial neural networks (ANN) to improve healthcare delivery at a reduced cost. Machine learning has been used in various instances to solve a series of real-life problems, especially in diagnosing diseases and the likes. In fact, a series of attempts have been made to use machine learning models to classify and identify diseases. However, the literature reveals that enough efforts have not been made to develop a system that can identify various types of workforce diseases or predict the probability of having such disease using machine learning models. The novelty of this research work, therefore, lies in the fact that, unlike the existing methods or techniques that are meant to identify a particular disease, the proposed system (not a technique or method) can identify many workforce diseases and also predict the possibility or probability of patients having such diseases. Hence, this research proposed a system to predict the likelihood of diseases among a cross-sectional workforce using machine learning models, improving the existing methods. Python programming language was used as an instrument to implement the system. The system has been tested and found to be highly efficient.

1.1 Literatures

Tejas and Pramila, (2018) presented research work on logistic regression and Support Vector Machine (SVM) based diabetics' prediction. The research work combined different machine learning techniques to predict diabetics at a very early stage. The steps adopted include data extraction, data preprocessing, and processing. The research work will go a long way to assist the medical doctors in detecting the disease at the initial stage. In their work, they considered, mainly, diabetics patients, whereas, in this research work, many workforce diseases will be considered.

Also, Yadar and Archad (2018) classified occupational stress using Machine Learning (ML) techniques. In their research work, a model was implemented to predict occupational stress in an industrial sector. The Machine Learning technique used includes Support Vector Machine (SVM), Neural Network (NN), Decision Tree(DT) and Random Forest(RF). Stratified cross-validation was used for training and testing. The accuracy, sensitivity, and specificity were 60%, 80%, and 60%, respectively.

For the experiment, a ten (10) fold cross-validation strategy was used. In the experiments, the performances of different classifiers were compared, and Random Forest algorithm performs better than others. This research restricted the number of classifiers tested to three (3), but six (6) classifiers will be considered in the proposed research Yue et al. used ML techniques to predict employee turnover and their performance compared with each other. The ML technique used includes decision tree, random forest, gradient boosting tree, an extreme gradient boosting method, logistic regression, SVM, NN, linear discriminant analysis, and K nearness neighbour. The SVM performed better than others. While Nida, Tim, and Whitney in 2019 also applied the principle of artificial neural networks in making organisational decisions. This research work was able to identify various characteristics of ANN in the area of healthcare and organisational decision making. . The research provides a broad understanding of the various applications of NN in the health organisation. The study concluded that ANN could be applied in solving a series of problems across all healthcare organisations. However, this research work focuses on decision-making in the health care sector but not workforce disease prediction. Again, the research was a review and not an actual implementation of any ML algorithm or technique.

Furthermore, Sandhya and Mahek (2019) presented research work that has to do with predicting mental disorder for employees in the IT industry. The researchers submitted that the employee should be regularly subjected to medical checkups to track their health status, especially concerning mental illness. The authors also submitted that stress would increase depression, heart attack, stroke, and cardiac arrest. Machine learning algorithms, including logistic regression, k-nearest neighbour, random forest, decision tree, bagging, boosting, and neural network with electroencephalogram (EFG) signal analysis, were used for the experiment. The data set was obtained from IT professionals from different locations and regions. Out of all the models, random forest appeared to be the best. It was recommended that employers and employees maintain a good relationship, creating a good environment for maximum productivity with "minimum tress". The research work was restricted to mental disorder, and other types of ailments were not considered.

Finally, Maria, Osei, and Md Al Mason in 2019 used a five ML algorithm to investigate the diagnosis and prognosis of cancer and heart disease. The algorithm used includes logistic regression, random forest, principal component regression, multivariate adaptive regression and SVM. The data used was split into train data set (67%) and test set (33%). The predictions' mean significance error, the misclassification rate, and the prediction accuracy (PA) were computed, and the values obtained were quite promising. The instrument used for the implementation was R programming language. The random forest and principal component regression were better in the analysis of breast cancer and heart disease.

The research work, unlike the proposed research, was restricted to only one disease.

Peter et al. (2020) presented a research work titled ML to predict primary care and advance workforce research. The researchers used (2014-2016) data. Three sets of random forest classifiers were used to predict speciality. F1 score and area under the receiver operating characteristics were used to evaluate the performance. The result obtained is better than the existing ones.

Sangiwe et al. (2018) predicted length of service among health care workers in undersaved communities in South Africa using an ML approach. The researcher used recruitment and retention data for Africa Health placement to develop the predictive model. A cross-validation technique was used to validate the model and compare the model in terms of its performance. R statistics programming language was used as an instrument of implementation. All the models form identical results. It was concluded that the ML models would serve as a useful tool in predicting the length of service for health workers.

Peter et al. (2020) used ML to predict primary care and advance workforce research. With the aid of medicare claim data, the authors trained three random forest classifiers to predict primary care and other specialities. The data set ranges from the year 2014 to 2016, and about 564 986 physicians were used. For evaluation, Area Under the Receiver Operating Characteristics (AUC_ROC) and F1 was used.

Going by the literature, there are series of related work, but none of the researchers attempted to develop a system that is capable of detecting workforce diseases using a machine learning approach or principles; rather, they were developing techniques or methods for disease identification using ML algorithms or models. So, this research work is an improvement on what we currently have on the ground, and this speaks a lot about the novelty of this research work.

2. Materials and methodologies

The steps used in the development of the machine learning models using both supervised and unsupervised learning are shown in Figure 1 below:

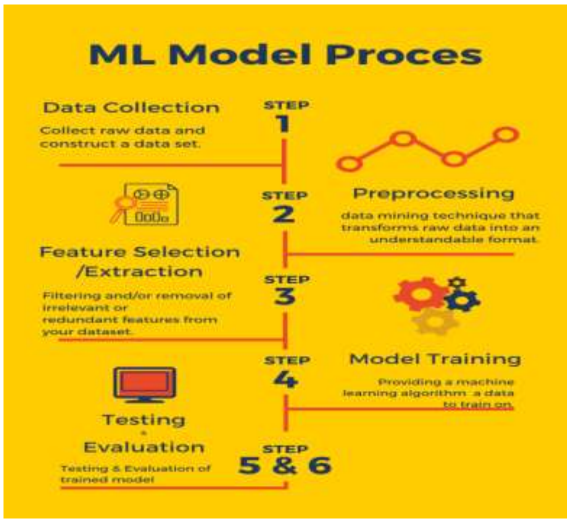


Figure I: Model Development Process

2.1. Data Collection

The data used to generate the datasets which the models used for training was obtained from different sources as explained as follows.

i) Data from research conducted in a rural community in Southwest Nigeria where we have a total of 1040 participants which comprises 330 males and 710 females having a mean age (± standard deviation) of 66.77 ± 12.06 years (range, 40–88 years). Most of the participants (56.7%) were between the ages of 30–60 years. Hypertension was present in 66.4%, diabetes mellitus in 4.8%, abdominal obesity in 38.46%, smoking in 2.9%, physical inactivity in 29.8%, and high alcohol consumption in 1%. Dyslipidaemia, as represented by low HDL-C, occurred in 30%. There were borderline high levels of TC in 4.5%, LDL-C in 1.1%, and TG in 12.5%, but no subject had a high level while abdominal obesity, alcohol consumption, and smoking were statistically significantly associated with sex.

ii) Also from Pima Indians Diabetes Database from the National Institute of Diabetes and Digestive and Kidney Diseases, India. All data patients were females with their age ranges from least 21 years old of Pima Indian heritage. Several constraints were placed on the selection of these instances, which includes the following: Glucose: Plasma glucose concentration within 2 hours, Pregnancies: Number of times pregnant, Blood Pressure: Diastolic blood pressure (mm Hg), Skin Thickness: Triceps skinfold thickness (mm), Insulin: 2-Hour serum insulin (mu U/ml), BMI: Body mass index (weight in kg/ (height in m) ^2), Diabetes Pedigree Function: Diabetes pedigree function, Age: Age (years), Outcome: Class variable (0 or 1)

iii) Cardiovascular Disease dataset from Kaggle. The dataset consists of 70 000 records of patients' data, 11 features + target. In this dataset, three types of input features were considered, namely Objective: information, examination: results of medical examination, and Subjective: the patient gave information. The Features are shown in Table 1 below.

Table I: Variable Features

Variable	Feature Type	Variable Name	Data Type
Age	Objective Feature	Age	int (days)
Height	Objective Feature	Height	int (cm)
Weight	Objective Feature	Weight	float (kg)
Gender	Objective Feature	Gender	categoric al code
Systolic blood pressure	Examinatio n Feature	ap_hi	Int
Diastolic blood pressure	Examinatio n Feature	ap_lo	Int
Cholesterol	Examinatio n Feature	Cholester ol	1: normal 2: above normal 3: well above normal
Glucose	Examinatio n Feature	Gluc	1: normal 2: above normal 3: well above normal
Smoking	Subjective Feature	Smoke	Binary
Alcohol intake	Subjective Feature	Alco	Binary
Physical activity	Subjective Feature	active	Binary
Presence or absence of cardiovascul ar disease	Target Variable	cardio	Binary

2.2 Pre-processing

This data mining technique was used to transform the raw dataset from the previous stage into an understandable format. The data imputation and data cleaning process ensured that the CSV files were properly labelled and that Nulls and hot encoding were removed. The data imputation was done using the imputation (mean/median) values to guess the values of missing datum/data independently.

2.3 Feature Selection/Extraction

At this stage, we used feature selection methods to reduce the number of attributes and then avoid redundant features. There are many feature selection methods. This study used PCA and minimum redundancy maximu m relevance (mRMR) to reduce dimensionality, reduce model complexity, and overfitting. Feature extraction and reduction are made using the Linear Discriminant Analysis (LDA) method.

2.4 Model Training

Methods used for training: The models were trained using supervised learning algorithms. The labels were binary (0 = Not Infected, 1 = Infected), and the probability of having the disease was also computed. Logistic regression Classifiers, Support Vector Machines, Nearest Neighbour classifiers, and Decision Trees were used to train the models. Multiple algorithms were used so that a comparison can be made in search of the choice of the best algorithm, which eventually produced the best model.

2.5 Model Testing

The models were tested thoroughly with test data. The K-Fold cross-validation was employed in testing the model. It resulted in a less biased or less optimistic estimate than other methods, such as a simple train/test split.

2.6 Model Evaluation

The models were evaluated using a confusion matrix. A confusion matrix is an N X N matrix, where N is the number of classes being predicted, and a tabular visualisation of the model predictions versus the ground-truth labels. Each row of the confusion matrix represents the instances in a predicted class, and each column represents the instances in an actual class. The results obtained during the evaluation process are displayed individually in figures 9, 10 and 11.

2.7 System Architecture

The system architecture (level1) used the various machine learning models to detect; Cardiovascular Diseases, Computer Vision Syndrome, Stress, Diabetes Mellitus), diseases, and the second level (level2) uses an insertion sort algorithm to return the disease with the highest probability to the user as shown in the architectural diagram in figure 2. The systemtakes input data and uses it to run all the machine learning models. The results are then saved in the database, sorted, and the most likely disease is displayed.

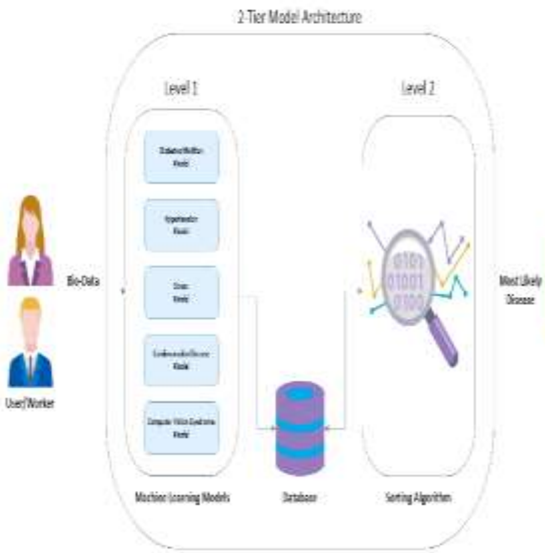


Figure 2: System Architecture.

2.8 Flowchart of the System

As shown in figure 3, the flowchart of the system shows the steps involved in its execution. It thoroughly shows the process flow from accepting data as input and running the model to display the result to the user. Each step in the process is represented by a different symbol and contains a short description of the process step. The flowchart symbols are linked together with arrows showing process flow direction.

Figure 3: System Flowchart

2.9 Use-Case diagram of the System

The use case diagram is a diagrammatic representation of how the user interacts with the proposed system. It shows all interactions or actions possible by the actors (Workers and Administrators) associated with the system, as explained in figure 4.

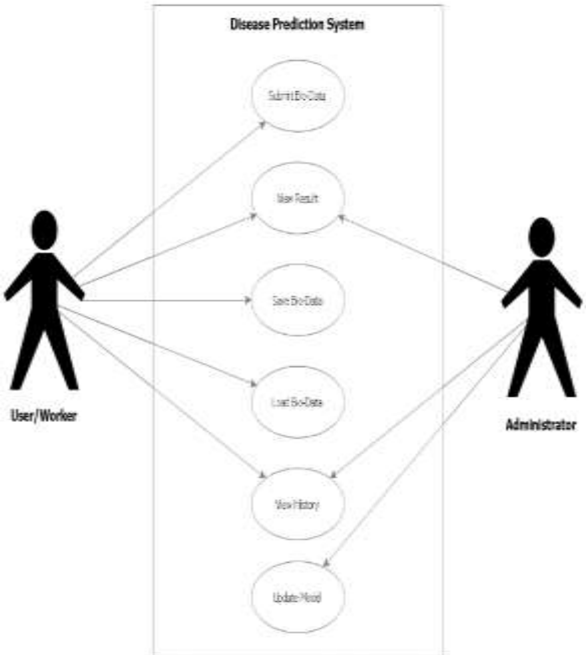


Figure 4: Use Case Diagram

3. Implementation and Results

The results obtained when the models were trained with various machine learning algorithms to identify the various workforce diseases were shown in figures 5, 6, 7, 8, 9, and 10. The implementation of the newly designed system was carried out, and the various system interfaces obtained at the implementation stages were displayed in figures 11, 12, 13, 14, 15, and 16.

3.1 Diabetes Model

Multiple machine learning algorithms were used to train data to obtain a model for the prediction of Diabetes. The models are Decision Tree, Hierarchical Clustering, K Means Clustering, Kernel SVM, KNN, Logistic Regression, Naïve Bayes, Random Forest, and SVM.

The models determine whether a person has Diabetes and also confirm its degree or extent. The prediction uses multiple data associated with the following features::Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Pregnancies. The models help to give a 2-class prediction ("0" Has Diabetes & "1" Does not Have Diabetes). In the model development process, the machine learning models were trained in PyCharm 2019 with Scikit-learn package. The level of accuracy and their confusion matrices are presented in figures 5. The Logistic Regression model had an accuracy score of 80.65%. The KNN model had an accuracy of 78.12%; the SVM model had an accuracy of 76.56%, the Naïve Bayes model had an accuracy of 80.08%, the Decision Tree model had an accuracy of 78.92%, and Random Forest has an accuracy of 77.00%

Figure 5: Diabetes Classification Algorithm and their Accuracies

Figure 6: CVD Classification Algorithm and their Accuracies

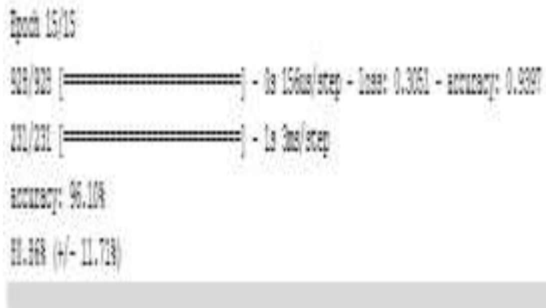
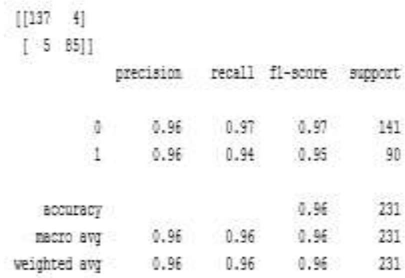


Figure 7: MLP model accuracy of Hypertension



[[137 4]
[5 85]]

	precision	recall	f1-score	support
0	0.96	0.97	0.97	141
1	0.96	0.94	0.95	90
accuracy			0.96	231
macro avg	0.96	0.96	0.96	231
weighted avg	0.96	0.96	0.96	231

Figure 8: Classification Report and Confusion Matrix of Hypertension's MLP model

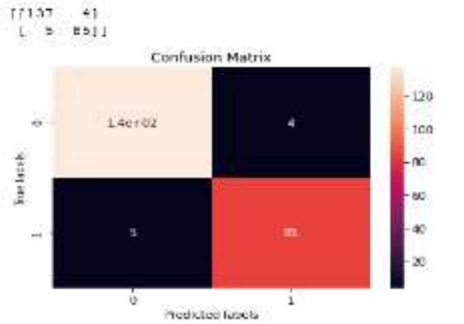


Figure 9: Confusion Matrix for the MLP Models

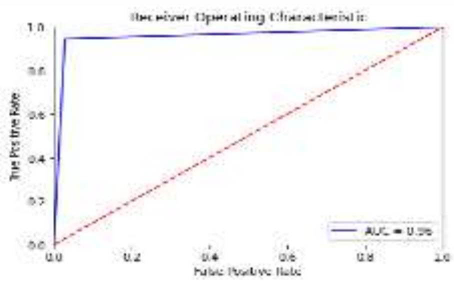


Figure 10: Area Under the Curve of MLP Model

Table 2: Stress Classification

Sr.No	Classifier	Accuracy
1	Guassian Naive Bayes	100%
2	K Nearest Neighbor	100%
3	DecisionTree	100%
4	Support Vector Machine(SVM)	100%
5	Neural Network(1 hidden layer)	96.67%

Figure 11: Home Page

Figure 12: The Main Page

Figure 13: The Disease Detection Page

Figure 14: The Disease Detection Page Contd.

Figure 15: The Disease Detection Page Contd.

Figure 16: The Result Page

3.2 Cardiovascular Disease Models

Multiple machine learning models were trained in predicting cardiovascular diseases. They are Decision Tree, Logistic Regression, Random Forest, and AdaBoost. The models determine whether a person has cardiovascular diseases and their extent. It uses multiple data to predict, and the data includes age, gender, height, weight, cholesterol level, glucose, smoke, and alcohol. The models produced 2- class prediction ("0" Has CVD & "1" Does not Have CVD). In the model development process, the machine learning models were trained in PyCharm 2019 with Scikit-learn package. Here are the models and their accuracy displayed in Fig 6. The Decision Tree Algorithm model had an accuracy of 72.4%; the Random forest algorithm model had an accuracy of 72.13%, the AdaBoost Algorithm model had an accuracy of 72.05%, the Random Tree Classifier model had an accuracy of 80.25%, the Decision Tree classifier model had an accuracy of 80.2%, the Logistic Regression model had an accuracy of 80.4%. The Random Forest model had an accuracy of 72.13.

3.3 Hypertension Models

Multiple machine learning models were trained in predicting Hypertension. They are the Decision Tree and a Multiple Layer Perceptron Neural Network. The models determine whether a person has cardiovascular diseases and their extent. It uses multiple data to predict. The data includes Sex, Body Weight, height, smoker, systolic blood pressure, diastolic blood pressure, max systolic blood pressure, and heart failure.

The models give 2 - class prediction ("0" Has Hypertension & "1" Does not Have Hypertension). In the model development process, the machine learning models were trained in PyCharm 2019 with package Scikit-learn, Keras, and TensorFlow.

Here is the accuracy of the MLP model with an accuracy of 96.10%, as displayed in Figure vii. Fig viii is the Classification Report for the MLP model, which shows the precision, recall, f1Score, and support values for the MLP model. Here is the Confusion Matrix for the MLP model of the hypertension algorithm, which shows the number of true negatives, false negatives, true positives, and false positives. We can deduce from this image in Figure ix that it has high accuracy. The Area Under the Receiver Operating Characteristic Curve (AUC ROC) for the MLP model is 0.96, which means a 96% chance that the model will distinguish between a healthy person and someone with Hypertension. As shown in figure x, the accuracy of the Decision Tree model is 99%, which is very high. It means that this model is as accurate as an actual Hypertension Decision Tree Model.

3.4 Stress Models

Multiple machine learning models were trained in predicting stress. They are Decision Tree, Gaussian, KNN, Neural Network, and SVM. The models determine whether a person has cardiovascular diseases and to which degree he/she does. It uses multiple data to predict, and they are listed as follows: ECG (mV), EMG (mV), Foot GSR (mV), Hand GSR (mV), HR (bpm) and RESP (mV). The models help give a 2-class prediction ("0" Has stress & "1" Does not Have stress). In the model development process, the nine machine learning models were trained in PyCharm 2019 with package Scikit-learn, Keras, and TensorFlow

Here are the models and their accuracies displayed in Table ii. The Gaussian Naïve Bayes, K Nearest Neighbour, Decision Tree, and Support Vector Machine models all have an accuracy of 100% in detecting stress, while the neural network has an accuracy of 96.67%. It appears to be the best model.

4. Discussion

The Disease Detection System has been developed on a desktop application. The application's graphical user interface was developed using the Kivy application designer package and was programmed in Python. The Disease Detector Software has the following pages:

4.1 The Home Page

The home page contains a splash screen introducing the application and an option to proceed to the disease detection page. The splash screen automatically changes to the disease prediction page after one (1) second. The diagram is displayed in figure 11. The purpose of this page is to aid aesthetic value that makes the application both branded and user-friendly.

4.2 The Main Page

The main page allows the user to select whether an option to either run all algorithms to detect the most likely disease or condition or detect the diseases individually. As shown in figure 12, click the 'Detect Most Likely Disease' button to proceed to another page that will run all the models and return the most likely disease. Clicking the 'Predict Individual Diseases' button will bring a pop-up that requires the user to select a disease and run the models specifically for the one chosen.

4.3 The Disease Detection Page

The disease detection page allows user input of the different biomedical requirements as input via textboxes. There is also a 'Predict' button that, when clicked, runs all the machine learning models, and then calculates their probabilities, ranks the outcomes, and then displays the results via a pop-up. This is shown in figure 13, 14, 15 and 16.

4.4 The Result Pop-Up Page

As shown in figure xvi, the results of the models are displayed in a pop-up that shows the most likely disease that a person has. It also goes further to display information on the performance of other models too

5. Conclusions

A systemhas been developed to detect multiple workforce diseases using machine learning models. It has proved to be more efficient than the existing systems. The system has been tested and proved to be more accurate in the detection of Diabetes, Stress, Hypertension, Cardiovascular Diseases, Stress, Computer Vision Syndrome, and other related diseases. It also displays the probabilities or possibilities of suffering from other types of diseases hierarchically. In conclusion, the aim of this study, which is to develop a system that predicts the likelihood of diseases among a cross-sectional workforce using machine learning models, was achieved.

References

Maria, C. Mariani, OseikTweneboah and Md Al Bhuiyan (2019). Supervised machine learning models applied to disease and diagnosis and prognosis. Public Health.www.aimspress.com/journal/aimsph

NidaShahid, Time Rappon and Whitney Berta (2019).Application of Artificial Neural Networks in Healthcare Organization decision making. A sceeping review PLOS/ONE 14(2) 0212356 http://doi.org/10.1371/journal.pone 2012356

Peter Wingrove, WistonLiaw, Jerremy Weis, Stephen Petterson, John Maier, Andrew Bazemores (2020) Using Machine Learning to Predict Primary Care and advance workforce. Annals of Family Medicine.

Yadar S. K. and ArshaidHashmi (2018).An investigation of occupational stress classification by using machine learning techniques.IC Journal International Journal of Computer Sciences and Engineering.www.ijcse@online.org

Sandhya, P. MaihekKaritesaria (2019). Prediction of mental disorder for employees in IT industry. International Journal of Innovative Technology and Exploring Engineering (IJITEF). ISSN: 2278-3075, Volume 8, Issue 65.

Sangiwe Moyo,Tuan Nguyen Doan, Jessica Ann YunandNdumisoTshuma(2018) , Application of Machine Learning Models in the in Predicting Length of Stay Among Healthcare workers in Undeserved Communities in South Africa, Human Resources for Heath.

Tejah N Joshi and Pramila, M. Chawan (2018). Logistic regression and SVM-based diabetics prediction system. International Journal for Technological Research in Engineering, Vol 5, Issue 11, ISSN (Online): 2347-4718 www.ANNFAMMED.ORG.vol.18, No4 July/August 2020

Yue Zhao, Maciej, K. Hrynetuicki, Francasca Cheng, Boyangfu, and Xiaoyu Zhu (2019).Employee turnover prediction with machine learning.A reliable approach. Research gate/328772915.

Maria, C. Mariani, OseikTweneboah and Md Al Bhuiyan (2019). Supervised machine learning models applied to disease and diagnosis and prognosis. Public Health.www.aimspress.com/journal/aimsph

NidaShahid, Time Rappon and Whitney Berta (2019).Application of Artificial Neural Networks in Healthcare Organization decision making. A sceeping review PLOS/ONE 14(2) 0212356 http://doi.org/10.1371/journal.pone 2012356

.