Capstone Project Concept Note and Implementation Plan

Project Title: [Chronic Diseases Prediction Using Machine Learning Techniques]

Team Members

1. Mariam Kili Bechir

Concept Note

1. Project Overview

Chronic diseases prediction is a vital subject because it addresses a critical need in healthcare which is a major concern for humanity. Detecting chronic diseases early is crucial, and that's where advanced automation systems such as AI and Machine learning come into play. As a part of Artificial Intelligence, Machine Learning is a powerful tool that can predict outcomes based on provided data. This study's goal is to enhance the healthcare systems and that's where it can be connected with the broader missions of Sustainable Development Goals (SDGs) that aim for the healthcare improvement. Embracing projects like the implementation of Machine Learning in multiple chronic diseases detection aligns seamlessly with the objectives of the United Nations SDGs. Recognizing healthcare as a cornerstone of sustainable development, it becomes apparent that not all healthcare systems worldwide are equally advanced. Particularly, individuals in regions lacking access to modern medical resources often grapple with challenges in the diagnosis and detection of chronic diseases.

In our study, we're diving into Machine Learning to predict common chronic diseases like diabetes, heart diseases, and Parkinson's the kind of health challenges that affect the metabolism, cardiovascular, and nervous systems, respectively and most people suffer from. This tech-driven approach holds the promise of improving how we spot these issues early, intervene effectively, contribute to SDGs and manage these health conditions more proactively.

2. Objectives

Chronic diseases pose a significant global health challenge, demanding early detection for effective intervention. Given the paramount importance of healthcare, this project addresses the critical need to enhance predictive capabilities for chronic diseases such as diabetes, heart diseases, and Parkinson's. By harnessing the power of AI and Machine Learning, we aspire to create a system that not only enhances accuracy but also expedites the diagnostic process of chronic diseases such as diabetes, heart disease and Parkinson disease. The implications are profound, potentially leading to early disease detection and more effective healthcare system.

Building chronic disease prediction using Machine Learning is essential because this subject not only include Machine Learning domain, but also healthcare field and Machine Learning project building pipeline. This project allows us to identify, preprocess and build successful

models using some Machine Learning methodologies. By critically assessing what has been done, we pave the way for innovation, ensuring that our research is not only novel but also informed by the collective wisdom of the scientific community. The project focuses on enhancing healthcare through early detection of chronic diseases like diabetes, heart diseases, and Parkinson's, aligning with United Nations Sustainable Development Goals (SDGs).

3. Background

Chronic diseases are long-term health conditions that have a significant impact on the quality of life and mortality of millions of people around the world. According to the World Health Organization (WHO), chronic diseases are responsible for 71% of all deaths globally, and they are projected to increase by 17% in the next decade [1] .

Common Chronic diseases include diabetes, heart disease, and Parkinson's, are influenced by genetic, environmental, and lifestyle factors. Existing solutions face challenges like limited accessibility and accuracy. While traditional healthcare approaches exist, the introduction of Machine Learning injects innovation into the process. A machine learning approach can offer several benefits and advantages over the existing solutions or initiatives related to chronic diseases. Potential applications include predicting disease risk, diagnosing based on symptoms and biomarkers, prognosing disease outcomes, and optimizing treatment. Machine learning can enhance accuracy and expedite diagnosis, leading to early disease detection. It also provides an accurate and timely results that can facilitate treatment decisions and provides a more effective healthcare system.

4. Methodology

Predicting chronic diseases is recognized as a classification problem. In our study, we are trying to classify the three different diseases separately. the outputs are 0 and 1 for each disease. 0 represents the person not affected by the disease and 1 is the result of the affected disease. It's the case for all of our three different datasets. Before building any model, we analyzed each dataset, use visualization techniques such as Matplotlib, Pandas Profiling and Seaborn to visualize our data and have a deep understanding about the different diseases. After doing Exploratory data analysis, To classify these diseases, Machine Learning classification algorithms such as Logistic Regression, Support Vector Machine, KNeareastNeighbors Classifiers and Bagging ensemble technique such as Random Forest and Boosting technique such as XGBoost Classifier are used to build three different models.

For heart disease dataset, the best model was found with Logistic Regression, the score obtained is 86% on the training data and 89% on the test dataset. 20% of our dataset is considered for evaluating and testing the model. These are the results obtained after using Hyper Parameter Tuning to optimize our model. The hyperparamter Tuning technique used is GridSearch. With this technique, we combined different algorithms together with their different hyperparameters and the best algorithms was Logistic Regression, RandomForest and Support Vector Machine (SVM).

	model	Best score	best_score
0	logistic_regression	0.834609	C= 0.23, max_iter= 100

1	random_forest	0.834609	max_depth=2, max_leaf_nodes= 6
3	SVM	0.822194	C= 100, kernel= 'linear'
4	xg_boost	0.797619	max_depth=9, max_leaf_nodes=3
5	KNN	0.702381	metric= 'manhattan', 'n_neighbors': 7

Table 1- sorted models results after hyper parameter tuning

For diabetes disease dataset, the algorithms used are Logistic Regression, SVM, RandomForest and XGboost. XGBoost gave the best score after Hyper Parameter Tuning. The score obtained by this algorithm is 96% on the training dataset and 77% on the test dataset. 20% of the dataset was used to test the model and 80% was used for training the data. We then used StandardScaler normalization technique to normalize our data. The training score with RandomForest gave us 91% and the test score with the same algorithm gave 71% as a result.

	model	**best_score**	**best_params**
0	svm	0.778529	{'C': 0.1, 'kernel': 'linear'}
1	knn	0.750860	{'n_neighbors': 7}
2	random_forest	0.780195	{'max_depth': 6, 'max_features': 'log2', 'n_es...
3	xg_boost	0.747608	{'n_estimators': 5}
4	logistic_regression	0.778529	{'C': 1, 'n_jobs': -1}

Table 2- models's results after hyper parameter tuning for diabetes disease

For Parkinson disease dataset, Dimensionality Reduction Technique known as (Principal Component analysis (PCA) was used to reduce the dimension of our features by keeping other's features importance.

For Parkinson disease dataset, the Machine Learning techniques used are: Logistic Regression, SVM, KNN, RandomForest and XGBoost. All these algorithms gave a good result but the best one obtained is with KNN. The scores obtained using KNN are 92 % on the training dataset and 90% on the test data. The second best algorithm for this data is SVM. The training score with SVM is: 85% and the test score with SVM is 87%. 80% of the data is used for training and 20% was used for testing and evaluation. GridSearch hyper Parameter tuning technique was used to perform the result and get this final score.

	model	best_score	best_params
0	svm	0.910081	{'C': 1, 'gamma': 1, 'kernel': 'rbf'}
1	knn	0.897581	{'metric': 'manhattan', 'n_neighbors': 11, 'we...
2	random_forest	0.878024	{'max_depth': 10, 'max_leaf_nodes': 9, 'n_esti...
3	xg_boost	0.884476	{'max_depth': 6, 'max_leaf_nodes': 3, 'n_estim...
4	logistic_regression	0.833266	{'C': 0.012742749857031334, 'max_iter': 100, '...

Table 3-Parkinson disease models results after hyper parameter tuning

To analyze the performance of our models, there are some metrics used such as confusion matrix, accuracy score and classification reports to help us evaluate our models.

After building, evaluating and the different models, an interface will be used to build to build a Multiple chronic disease prediction system. For this, we will use Streamlit to build the web application and then deploy the application built.

5. Architecture Design Diagram

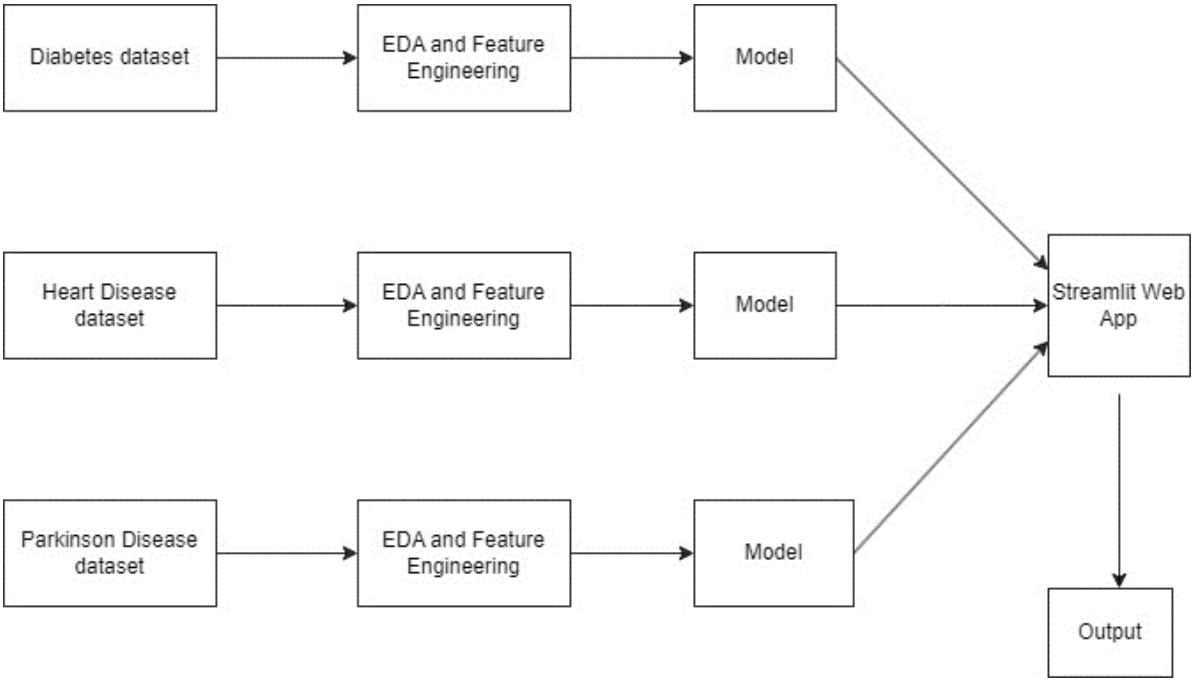


Figure 1- Project Design

This figure represents our project architecture design. After introduced the 3 datasets of the three different chronic diseases, we analyzed the data and used some feature engineering and extract some insights and making visualization. After that, we trained each dataset with different Machine Learning algorithms and we got three different models. These models are then used to predict our next diseases via an application built using Streamlit.

6. Data Sources

The data used for this study are collected from Kaggle which is an open-source platform for datasets. It's a collection of three different datasets. Each dataset represents 1 type of chronic diseases. Totally we have 3 different chronic diseases which are: diabetes, heart disease and Parkinson diseases. Each dataset size is different from the other. The total of diabetes dataset is 768 entries and 8 independent features which are: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and the Outcome feature which is the target that will going to be trained as y_train and predicted later. For heart disease, the total entries are 303 rows and we have 14 attributes in which one attribute is the target feature that need to be predicted. About Parkinson disease, 195 entries are registred with 24 different features in which 23 are independent and 1 feature represent the target dependent feature. To analyse these data, extract insights and make decision about these data, Exploratory Data Analysis (EDA) techniques are used. The data don't have missing values and outliers. We analysing statistically each datasets, visualize data and make insights etc.

7. Literature Review

According to existing research's, building chronic disease prediction using Machine Learning is essential because this subject not only include Machine Learning domain, but also healthcare field and Machine Learning project building pipeline. This project allows us to identify, preprocess and build successful models using some methodologies. By critically assessing what has been done, we pave the way for innovation, ensuring that our research is not only novel but also informed by the collective wisdom of the scientific community.

- In their conference published in 2016 [2] In their conference published in 2016, (Astya et al., n.d.) discuss about methods used to predict diagnostic codes for chronic diseases using machine learning techniques. They focused on 11 types of different chronic diseases such as kidney disease, osteoporosis, arthritis etc and used Machine learning techniques for each disease's diagnosis.

- [3] In this paper published in 2018, (IEEE Circuits and Systems Society. India Chapter and Institute of Electrical and Electronics Engineers n.d.) five chronicle datasets are taken and the machine learning algorithms such as decision tree, random forest, and the support vector machine are applied and the predicted whether the patient is suffering from a chronicle disease such as heart disease, liver disease or diabetes. A result was obtained by comparing all algorithms performance on all dataset the random forest predicts with high accuracy.
- [4] use Machine learning predictive models to predict chronic diseases. They precise that among the methods considered, support vector machines (SVM), logistic regression (LR), clustering were the most commonly used. These models are highly applicable in classification, and diagnosis of CD and are expected to become more important in medical practice in the near future.
- Logistic regression was as good as machine learning for predicting major chronic diseases [5] to evaluate the performance of machine learning (ML) algorithms and to compare them with logistic regression for the prediction of risk of cardiovascular diseases (CVDs), chronic kidney disease (CKD), diabetes (DM), and hypertension (HTN) .
- [6] demonstrates the noncrisp Rough K-means (RKM) clustering for figuring out the ambiguity in chronic disease dataset to improve the performance of the system. The

experimental results demonstrate that the proposed system is successfully employed for the diagnosis of chronic diseases. The proposed model achieved the best results with naive Bayes with RKM for the classification of diabetic disease (80.55%), whereas SVM with RKM for the classification of kidney disease achieved 100% and SVM with RKM for the classification of cancer disease achieved 97.53 with respect to accuracy metric.

- In [7] study, a comparison between deep learning model and twelve machine learning and ensemble learning methods based on relatively small data including 183 healthy individuals and 401 early Parkinson Disease patients shows the high detection performance of the designed model, with 96.45% of accuracy.
- [8] combines Deep Learning and Machine Learning techniques to predict heart disease. Using deep learning approach, 94.2% accuracy was obtained.
- In this study [9] Naïve Bayes and KNN algorithms was used. The accuracy of heart disease prediction using naive Bayes obtained was 94.5% which is greater than accuracy of KNN. They then compared naive Bayes with KNN and figure out that KNN requires more memory and time. Additionally, a risk prediction system using the CNN algorithm was developed to assess the risk of heart disease.
- [10] proposed a study that concerns the cardiac disease diagnosis. After applying to preprocess and feature engineering techniques, machine learning approaches like random forest, decision trees, gradient boosted trees, linear support vector classifier, logistic regression, one-vs-rest, and multilayer perceptron are used to perform binary and multiclassification on the dataset.
- In the study [11] published in 2022, a Deep learning model for multi-classification of infectious diseases was build based on unstructured electronic medical records in order to assist in clinical infectious-disease decision-making. The accuracy of MIDDM (Multi-classification of Infectious diseases Model) achieved is 99.44%, which is significantly higher than that of XGBoost (96.19%), Decision tree (90.13%), Bayesian method (85.19%), and logistic regression (91.26%).
- In this study, [12] various machine learning algorithms have been used in the training process to predict diseases belonging to different branches of medicine, such as diabetes, bronchial asthma, and covid. It is also the first study to achieve an accuracy score of 99.33% with a dataset that involves a greater number of diseases.

From this literature review, we can conclude that researches based on the application of Machine Learning techniques in the healthcare domain continue to grow from each year to other. The analyses of these studies from 2016 to 2023 show us that we have a continuous grow in the applied Machine Learning and artificial intelligence techniques in the healthcare domain, especially in diseases diagnosis such as chronic diseases. This leads to a conclusion where we can say Machine Learning techniques including Deep Learning have a huge contribution in the development of healthcare.

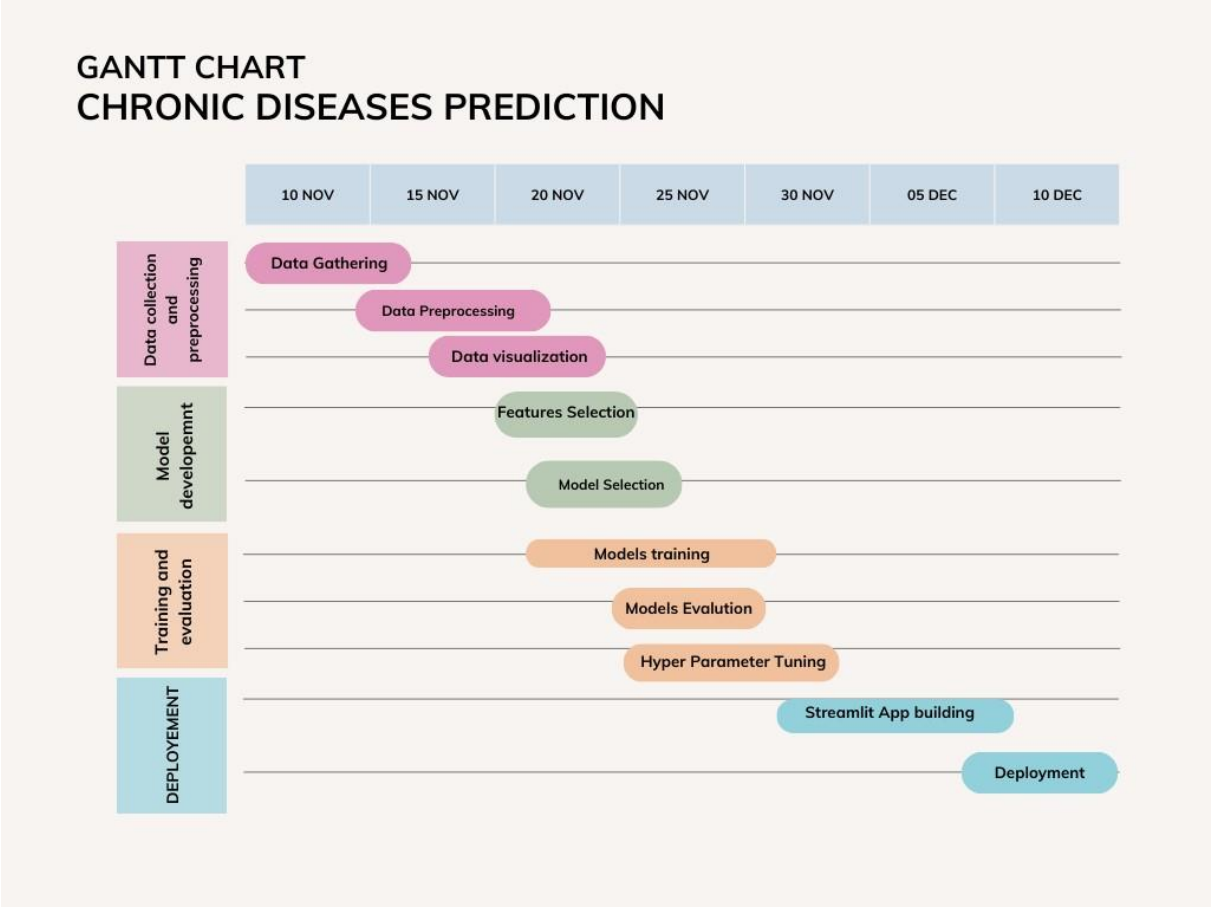
Implementation Plan

1. Technology Stack

The project is being implemented using the following technologies and tools:programming language: Python

- Libraries used in this project: Pandas, Numpy, Scikit Learn, Matplotlib, Seaborn : Empowering the project with the data manipulation capabilities of Pandas, numerical operations with Numpy, and the comprehensive tools for machine learning and visualization provided by Scikit Learn, Matplotlib, and Seaborn.
- Frameworks: Streamlit is used for creating an interactive and user-friendly interface with the Streamlit framework to showcase and explore the predictive models and results seamlessly.
- Softwares used: Jupyter Notebook and VScode are used.

3. Timeline



3. Milestones

After Building and evaluating the our models, here are the classification metrics that show us the entire performance of our classification models. These classification report metric is important to check because it gives us a complete result about our data and the performances our model trained like Precison, recall, f1-score and accuracy.

	precision	recall	f1-score	support
0	0.80	0.86	0.83	100
1	0.70	0.61	0.65	54
accuracy			0.77	154
macro avg	0.75	0.74	0.74	154
weighted avg	0.77	0.77	0.77	154

Table 4-Classification_report for Diabetes data with XGBoost

	precision	recall	f1-score	support
0	0.89	0.86	0.88	29
1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

Table 5- Classification_report for heart disease with LogistciRegression

	precision	recall	f1-score	support
0	0.71	0.71	0.71	7
1	0.94	0.94	0.94	32
accuracy			0.90	39
macro avg	0.83	0.88	0.83	39
weighted avg	0.89	0.89	0.90	39

Table 6- Classification_report for Parkinson disease with KNN

4. Challenges and Mitigations

Looking at the model performances of diabetes disease dataset, the accuracy is 0.77 for the test data, and the recall and f1-score are not really high. They are 0.61 and 0.65 respectively for the class 1. This is most probability due to the imbalanced dataset that we have between 0 and 1 classes. This may lead to some wrongs predictions especially False negative results would need to be decreased in this case.

The result of the model performance of heart disease dataset looks good when using Logistic Regression and we see the Precision, recall, f1-score and accuracy results for both classes 0 and 1.

When we look at the Parkinson model built, we see that the model performed well, especially for the class 1 but we might got some wrongs results because of the result of the performances the class 0. That might not really affect our model result. The accuracy also is a good accuracy 0.90 on the test data.

5. Ethical Considerations

In our study, we used a public dataset already shared on Kaggle. These data are used in many studies and the ethical aspect should be considered. The ethical aspects, such as privacy concerns in health data, is important to be considered. Researching and implementing robust ethical frameworks for handling sensitive health data is vital for responsible AI in healthcare.

6. References

[1] R. Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach," *J Healthc Eng*, vol. 2022, 2022, doi: 10.1155/2022/2826127.

[2] P. N. Astya, Galgotias University. School of Computing Science and Engineering, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section. SP/C Joint Chapter, and Institute of Electrical and Electronics Engineers, *Proceeding, International Conference on Computing, Communication and Automation (ICCCA 2016) : 29-30 April, 2016*.

[3] IEEE Circuits and Systems Society. India Chapter and Institute of Electrical and Electronics Engineers, *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET).*

[4] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," *J Pers Med*, vol. 10, no. 2, Jun. 2020, doi: 10.3390/jpm10020021.

[5] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J Clin Epidemiol*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/J.JCLINEPI.2020.03.002.

[6] T. H. H. Aldhyani, A. S. Alshebami, and M. Y. Alzahrani, "Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms," *J Healthc Eng*, vol. 2020, 2020, doi: 10.1155/2020/4984967.

[7] W. Wang, J. Lee, F. Harrou, and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning," *IEEE Access*, vol. 8, pp. 147635–147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

[8] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/8387680.

[9] S. Vilas and A. M. S. Scholar, "Diseases Prediction Model using Machine Learning Technique", doi: 10.32628/IJSRST.

[10] D. Hamid, S. S. Ullah, J. Iqbal, S. Hussain, C. A. U. Hassan, and F. Umar, "A Machine Learning in Binary and Multiclassification Results on Imbalanced Heart Disease Data Stream," *J Sens*, vol. 2022, 2022, doi: 10.1155/2022/8400622.

[11] M. Wang, Z. Wei, M. Jia, L. Chen, and H. Ji, "Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records," *BMC Med Inform Decis Mak*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12911-022-01776-y.

[12] M. ÇOLAK, T. TÜMER SİVRİ, N. PERVAN AKMAN, A. BERKOL, and Y. EKİCİ, "Disease prognosis using machine learning algorithms based on new clinical dataset," *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, vol. 65, no. 1, pp. 52–68, Jun. 2023, doi: 10.33769/aupse.1215962.