**Predictive Models for Breast Cancer Diagnosis using Machine Learning Algorithms**

**NAME: JAYDEEP PUROHIT**

**REGISTRATION NUMBER: 2202140**

A thesis submitted for the degree of master of science in Electronic

Supervisor : Prof. Dongbing GU

School Of Computer Science and Electronic Engineering

University of essex

August 2023

Contents

[List of Figures and Description 3](#_Toc143855237)

[List of Tables and Description 4](#_Toc143855238)

[ABSTRACT 5](#_Toc143855239)

[Introduction 6](#_Toc143855240)

[Problem Statement 9](#_Toc143855241)

[Objectives 9](#_Toc143855242)

[Related Work 10](#_Toc143855243)

[3. Dataset and Evaluation Metric 13](#_Toc143855244)

[3.1 Description of Dataset 13](#_Toc143855245)

[1). WDBC (Wisconsin Diagnostic Breast Cancer) 13](#_Toc143855246)

[3.2 Performance Evaluation Matrices 14](#_Toc143855247)

[4. Proposed Methodology 15](#_Toc143855248)

[4.1 Proposed Framework 15](#_Toc143855249)

[4.2 DET (Data Exploratory Techniques) 16](#_Toc143855250)

[4.3 Machine Learning Predictive Algorithms 17](#_Toc143855251)

[4.4. Experimental Setup 19](#_Toc143855252)

[Results and Discussions 21](#_Toc143855253)

[5.1 EDA: 21](#_Toc143855254)

[5.2 Machine Learning Predictive Algorithms Evaluation 23](#_Toc143855255)

[5.2.1. Support Vectoe Machine (SVM): 23](#_Toc143855256)

[5.2.2. Linear Regression (LR): 24](#_Toc143855257)

[5.2.3 K-Nearest Neighbor (KNN): 25](#_Toc143855258)

[5.2.4 Ensemble Classifiers (EC): 25](#_Toc143855259)

[5.3 Analysis of Classifier's Comparability 26](#_Toc143855260)

[5.4 Compared to earlier studies 27](#_Toc143855261)

[6.Project Plan and Management 29](#_Toc143855262)

[7. Discussion 30](#_Toc143855263)

[8. Conclusion 32](#_Toc143855264)

[9. References 33](#_Toc143855265)

# List of Figures and Description

|  |  |
| --- | --- |
| **Sr No.** | **Figure Description** |
| Figure 1 | Proposed diagram of BC using ML algorithms |
| Figure 2 | The BC dataset's class distributions show that WDBC is classified as benign and malignant |
| Figure 3 | Distribution of features in the WDBC dataset using various samples |
| Figure 4 | Feature correlation between several data samples such as Perimeter, Symmetry as well as Smoothness |
| Figure 5 | SVM-kernels' performance is compared via cross-validation |
| Figure 6 | Comparisons of the cross-validation scores and learning curves for (a) LR (b) LR with RFE. |
| Figure 7 | Performance comparison between KNN with hyperparameter and KNN with basic KNN |
| Figure 8 | All prediction models and techniques are compared using cross-validation matrices |

# List of Tables and Description

|  |  |
| --- | --- |
| **Sr No.** | **Table Description** |
| Table 1 | WDBC dataset features classification. |
| Table 2 | Testbed of Implementations |
| Table 3 | Comparison of the performance of linear and polynomial SVM kernels using the WDBC training and testing datasets |
| Table 4 | LR, LR-recall predication, LR-RFE as well as performance with basic LR-cross-validation |
| Table 5 | Voting classifier-recall, E-LR, and voting classifier (CV) performance comparison |
| Table 6 | Comparison of the execution times for each model and the highest level of accuracy |
| Table 7 | Comparison of our suggested BC prediction models' accuracy with that of earlier research that made use of the same WDBC dataset. |
| Table 8 | Project Plan and Management |

# ABSTRACT

Among American women, breast cancer has the highest mortality rate. Early detection methods for breast cancer diagnosis are promised by machine learning (ML) based predictive models. Making an assessment for models that accurately identify cancer is still difficult, though. To increase the accuracy of the diagnosis of breast cancer, we built four alternative predictive models and proposed DET in this study. In order to find the reliable feature categorization into classes of malignant as well as benign before models, four-layered essential DET, such as feature distribution, correlation, removal, and hyperparameter optimization, were thoroughly investigated. On the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, these suggested methods and classifiers were put into practice. The effectiveness and training duration of each classifier were evaluated using common performance indicators, such as confusion matrices as well as K-fold CV methods. With our DET, the models' diagnostic abilities increased; specifically, polynomial SVM gained 99.3% accuracy, LR had 98.06% accuracy, KNN had 97.35% accuracy, and EC had accuracy through the WDBC dataset of 97.61%. We also assessed the correctness of our key findings in comparison to earlier research. The process of implementation and the results can direct doctors towards adopting a useful model for a realistic knowledge of and outlook for BC tumors.

# Introduction

Following lung cancer, breast cancer (BC) is the largest cause of mortality for women worldwide, accounting for roughly 2.2 million new cases as well as 0.6 million new fatalities in 2020 (Sung, H., et al., 2021). 43,600 female fatalities from breast cancer were reported in the United States in 2021, while 281,550 new cases were diagnosed there (Siegel, R.L., et al., 2022). Breast cancer often develops in BC, more specifically in the inside layer of the milk-supplying lobules that provide milk to the milk conduit. Cancer cells arise from healthy cells as a result of DNA and RNA alterations or mutations (Natalie Baughan, BS, et al., 2022). The entropy rise may cause these alterations or mutations to happen spontaneously, or they may be brought on by other sources. For instance, radiation from nuclear sources, bacteria, viruses, fungi, and parasites, chemicals in the air, heat, food, and water, free radicals, mechanical cell-level injury, evolution, and ageing of DNA and RNA, as well as electromagnetic radiation such as X-rays, oven, gamma, and ultraviolet rays, etc., as well as nuclear radiation (Leo, et al., 2021). Tumors fall into two categories: benign and malignant, respectively. Though benign is not deadly and does not pose a threat to life, it may increase the risk of breast cancer. Malignant tumors, on the other hand, are more concerning and cancerous. According to a study that detected BC, 20% of women died from malignant tumors (Subashini, T.S., et al., 2009).

This research provides a strong emphasis on tumor diagnosis, which is now a hot topic in biomedicine. For the purpose of predicting breast cancer, the researchers are using data mining as well as machine learning technologies (Abdar, M., et al., 2020). On DM and ML, classifier-based prediction models can reduce diagnosis errors and improve the effectiveness of a cancer diagnosis. DM is a comprehensive amalgamation of several ways to unearth buried information and knowledge from massive datasets that are challenging to directly analyze. It has been extensively employed in the deployment of the prediction system for several diseases, including thyroid cancer (Park, K.H., et al., 2021) and lung cancer (McWilliam, A., et al., 2016). Fuzzy genetics and computer-aided systems have both been used to diagnose breast cancer using DM and ML techniques (Park, E.Y., et al., 2021; Bicchierai, G., et al., 2021). By evaluating the classifier and forecasting the approaching tumor based on historical data, the outcomes of these investigations correctly put the characteristics into two categories of tumors.

A study published in the literature demonstrated that BC prediction using ML classifiers in the early stages not only increases the chances of survival but also has the power to stop the spread of cancerous cells throughout the body (Akay, M.F., 2009). For instance, a study (Furey, T.S., et al., 2000) used the SVM-based technique for BC diagnosis and produced useful results in prediction. Like this, Furey et al. (2015) used SVM with a linear kernel for cancer tissue categorization and got a 93.4% accuracy. Later, Zheng et al. (2014) expanded on this work by providing a K-SVM hybrid model with 97% accuracy for classifying the WDBC dataset (Seddik, A.F.; Shawky, D.M., 2015). While Seddik et al. (2015) proposed a strategy based on tumor factors to get good WDBC data regarding a binary logistic version to identify breast cancer results, other researchers were working on various classifiers. Like this, Mert et al. developed a feature reduction technique independently generated component analysis to detect BC using a KNN classifier. It computed performance, distributed the features with a reduced one feature (1C) and 30 features, and achieved 91% accuracy (Abdar, M., et al., 2018).

This research described above have not considered the data exploration strategies, which allow the data mining methods to be more reliable and effective performance, aside from these favorable accuracies using various classifiers and methods. Numerous research (Rajaguru, H., 2019) encounter the accuracy restriction of ML classifiers due to the lack of such crucial methodologies. Due to incorrectly predicted true negative as well as false negative matrices, the confusion matrices in those experiments misidentified the malignant and benign classifications. In that earlier research that employed criteria to evaluate the benefit of nonlinear training classification, another flaw was discovered. However, as features are added, the performance of the model's execution time quickly improves (Brause, R.W., 2001). Considering this, the prediction model slows down, which compromises the precision of the diagnosis. For the data analyst and doctor, however, the model's accuracy and complexity of time are crucial concerns. We proposed DM approaches with various ML models to conduct a novel study for the diagnosis of breast cancer in response to the challenges and conclusions previously described (Yash Amethiya, et al., 2022).

Four alternative prediction models were developed in this study using four machine learning algorithms such as SVM, KNN, LR and EC to deal with a large volume of tumor features and extract key data for the detection of breast cancer (Sammut, SJ., *et al.,* 2022). We discovered that certain Literature lacks precision limits in the absence of these methodologies (Xueheng Liang, Xingyan Yu, Tianhu Gao, 2022). Though image data remain better appropriate for detecting breast cancer, we have not taken them into account in our study because we are using the WDBC dataset to use sophisticated ML classifiers. It provides a platform for investigating the implementation strategy for breast cancer detection by fusing DET and predictive models.

The tumor characteristics can be described in detail, which results in redundant data (V. Durga Prasad Jasti, et al., 2022). Because of the lengthy processing periods, such characteristics produce tedious results. Therefore, our primary objective was to examine a time-complex predictive model for cancer diagnosis in addition to a successful predicting model with realizable accuracy. Our models will be able to extract and mine important data from a large dataset by identifying correlations and removing the characteristics thanks to the consideration of time efficiency (Epimack Michael, et al., 2022).

The results showed adequate accuracy for the diagnosis of breast cancer with the shortest computation time, indicating the superiority of our study over others. This effort will make it possible for a data analyst to analyze breast cancer data using an intelligent ML model. Like this, a doctor would accurately identify breast cancer based on the classification of the tumor (Sushovan Chaudhury, et al., 2022).

Specifically, this study has made the following notable contributions: (1). With the WDBC BC dataset, which improved in quality by identifying the tumor and classifying it as benign or malignant, we looked at 4-prediction ML algorithms such as SVM, KNN, LR and EC (2). Before putting several ML classifiers into use as prediction models, these methods the forecasting methods to reach their highest levels of diagnosis accuracy for BC (3). We design tests to verify the models' time complexity prediction and classification accuracy, and we provide comparative analysis with cutting-edge studies and different evaluation matrices.

# Problem Statement

With more features, the BC algorithm performs better more quickly. The accuracy of the diagnosis is impacted by the prediction model's slowdown. For the data analyst and doctor, however, the model's accuracy as well as time complexity remain crucial concerns. A strategy to separating cancer into usual as well as tumor kinds was recently demonstrated in a study published in the journal Nature Cancer. While some of them performed trials using just one classifier, several research have employed the SVM classifier to predict breast cancer. Though, there remains still a need to investigate effective classifiers for predicting breast cancer using more efficient techniques like SVM, LR, KNN, and EC. In order to diagnose breast cancer, this study used 4 several prediction algorithms and suitable data mining and data exploration tools.

# Objectives

1. With the WDBC BC dataset, which improved in quality by identifying the tumor and classifying it as benign or malignant, we looked at 4-prediction ML algorithms such as SVM, KNN, LR as well as EC.
2. Before putting several ML classifiers into use as prediction models, these methods the forecasting methods to reach their highest levels of diagnosis accuracy for BC.
3. We design tests to verify the models' time complexity prediction as well as classification accuracy, and we provide comparative analysis with literature review as well as different evaluation matrices.

# Related Work

Worldwide, BC illness is a major cause of death. The most recent technology equips professionals with a variety of adaptive approaches to find BC in women after the classic cancer screening procedures (Rabiei R, et al., 2022). Numerous data science strategies work in conjunction with modern technologies to help gather and analyze data on cancer in order to forecast this deadly disease. Among these DS technologies, ML algorithms have been effectively used for cancer-based data analysis. For instance, studies were done to demonstrate that these ML algorithms could increase diagnostic accuracy (Huang, C.L., et al., 2008). It appears out that a skilled doctor attained a diagnosis accuracy of 79.97%. With machine learning, 91.1% of predictions were accurate.

The use of ML in the medical industry has gradually grown over the past two decades. However, the information gathered from the patients and assessment by the medical professional are the crucial elements for diagnosis. The ML classifiers can help to reduce human error and have quickly and thoroughly analyzed medical data (Polat, K.; Güne's, 2007). In our work, we used SVM, LR, KNN and EC for BC prediction among other ML classifiers for data modelling and prediction.

Due to SVM's best prediction accuracy, it was an extensively used ML technique in the detection of BC. According to Zheng, B., et al. (2014), Furey et al. (2000) demonstrated SVM for cancer tissue using a linear kernel identification and attained tolerable accuracy. Like this, Polat et al. (2007) achieved a 98.53% accuracy using SVM with least squares for predicting BC by removing duplicate characteristics.

According to Prasad, Y., et al. (2010), least square SVM proved helpful for training models using linear equations. His approach, nonetheless, was unable to fulfil the feature selection procedure. In order to integrate several technologies (Muzammal, M.; Qu, Q.; Nasrulin, B., 2019) provided a distributed database for multi-active features. A heuristic methodology for training the SVM classifier using a feature subset was put out by Prasad, Jain, and colleagues in 2010. It accurately divides the data on breast cancer into two groups with 91.7%. However, if the author uses the process for eliminating a feature to remove the noise data, this accuracy can be sufficiently improved (Jiang, X.; Xu, C., 2022).

In a similar vein, K-mean as well as SVM classifiers were combined in a hybrid model proposed by Zheng et al. (2014). By using the selection and extraction of features method, this model's goal was to identify the tumor features from the WDBC dataset. To differentiate between benign and malignant tumor patterns, a K-mean classifier was used (Kirola M, et al., 2022). The resulting patterns are calculated and taken into consideration as fresh patterns for the SVM model's training. Then, SVM is used to anticipate the arrival of tumors. Their hybrid model was used, which increased accuracy to 97%. The core tasks for data preparation, however—data exploratory techniques—have not been sufficiently addressed in order to train the suggested model (Seddik, A.F.; Shawky, D.M., 2015).

In addition to the SVM used optimal parameters for logistic regression (LR) on the WDBC datasets (Lim and Sohn et al., 2013). With optimized feature sets, it achieved 97.8% accuracy that was suitable for the WDBC dataset (Yang, L.; Xu, Z., 2019).

Seddik et al. (2015) proposed a binary logistic model based on variables with tumor imaging features for the identification of BC data. The suggested model correctly classified the WDBC data into benign and malignant categories with an average classification accuracy of 98% with the different features such as Area, texture, concavity, and symmetry were identified as important WDBC properties by this regression model (Mert, A., et al., 2015).

In order to predict breast cancer, Mert A. et al. (2015) presented a feature reduction strategy with independent component analysis. It effectively classified the WDBC features using 30 features and one reduced feature (1C) features by using the KNN classifier. With a 91% accuracy rate, it calculated performance using various matrices (Abdar, M., et al., 2018).

Rajaguru et al. (2019), who took on the BC prediction problem by applying the KNN and DT algorithms to categorize the WDBC features, later enhanced this study. KNN outperformed the DT in the feature classification using a conventional PCA feature selection method, it was discovered (Mushtaq, Z., et al., 2020). KNN achieved 96.4% accuracy with PCA method in a different study by Yang and Xu et al. (2019) (Fu, Y., et al., 2020). The usefulness of KNN has recently been investigated using two dissimilar BC datasets, considering its efficiency by the k values and several distance functions. It entails the three various experiment types: KNN, SVM and Chi-squares. the third strategy, Chi-square-based feature selection, had the highest accuracy (Kamyab, M., et al., 2018).

Few studies consider the EC with voting mechanism, the fourth prediction model, when making BC predictions. To distinguish benign tumors from malignant BC, M. Abdar et al. (2020) suggested an ensemble method via vote/voting classifier. For ML algorithms, it created a two-layer voting classifier. The outcomes of these voting methods showed that the straightforward classification algorithm performed as expected (Abdar, M., 2020). We were inspired to conduct experiments using voting classifiers and various ML approaches as a result of these investigations. The extent possible knowledge, none of the methods has made use of feature correlation and removal for the specific BC dataset.

This research carried out tests to categorize the characteristics of cancer, which remains a difficult problem. A technique to separating becoming normal and tumor tissues was recently demonstrated in a paper published in Nature Cancer (Nusrat Mohamud din, et al., 2022). The SVM classifier has been employed in numerous research for BC prediction, but just one classifier was used in some of the tests. There is still a need to investigate the effective classifier for BC prediction, according to studies by Abdar, 2020; Seddik, A.F., et al.; Mert, A., et al.; and Mushtaq, Z., et al. In order to diagnose BC, this study used four distinct forecasting models and suitable data mining exploratory tools.

# 3. Dataset and Evaluation Metric

## 3.1 Description of Dataset

The experiments in this study were run on WDBC datasets. These datasets were chosen since they have been extensively utilized in several publications (Abdar, M., et al., 2018) (Durgalakshmi, B.; Vijayakumar, V., 2020) (R. R. Sekar, et al., 2023) (M. López-Pérez; et al., 2023). The ML algorithms that produce sufficient accuracy by dataset remained also trained.

## 1). WDBC (Wisconsin Diagnostic Breast Cancer)

The WDBC dataset contains data from 569 patients and includes 10 breast tumor feature characteristics. It was delivered by Dr. William H. Wolberg at the University of Wisconsin Madison's General Surgery Department in the USA. The dataset link is (<https://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/WDBC/>) Fluid samples from patients' solid breast lumps were used to produce this dataset. Or WDBC dataset available on Kaggle website and dataset source link is here: (<https://www.kaggle.com/code/meetnagadia/breast-cancer-classification/input>) Based on the digital scan, cytological feature analysis was then carried out using software named Xcyt. Finally, the diagnosis of each sample such as Malignant (M) or Benign (B). In summary, the dataset had 569 occurrences and 32 attributes.

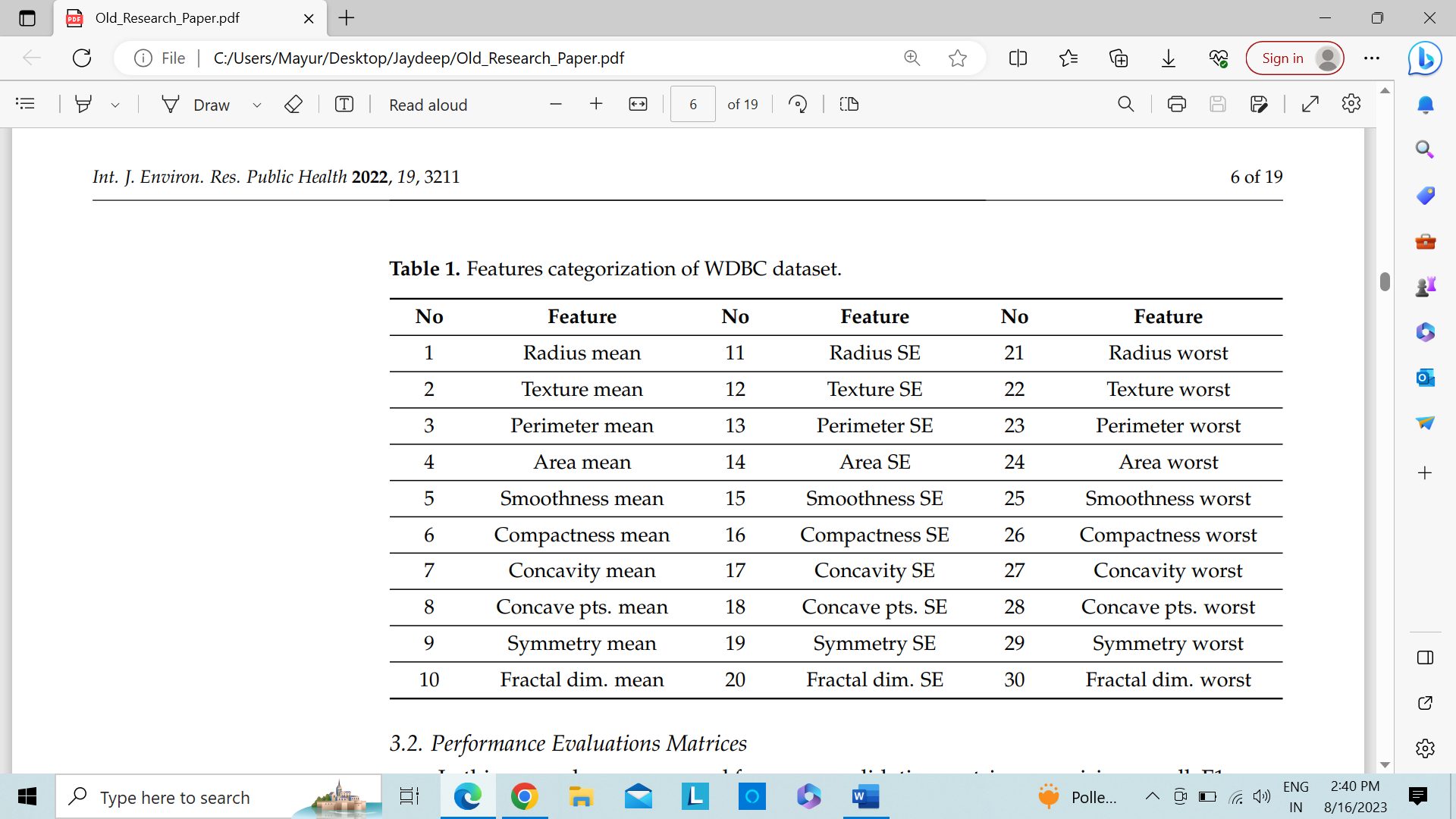
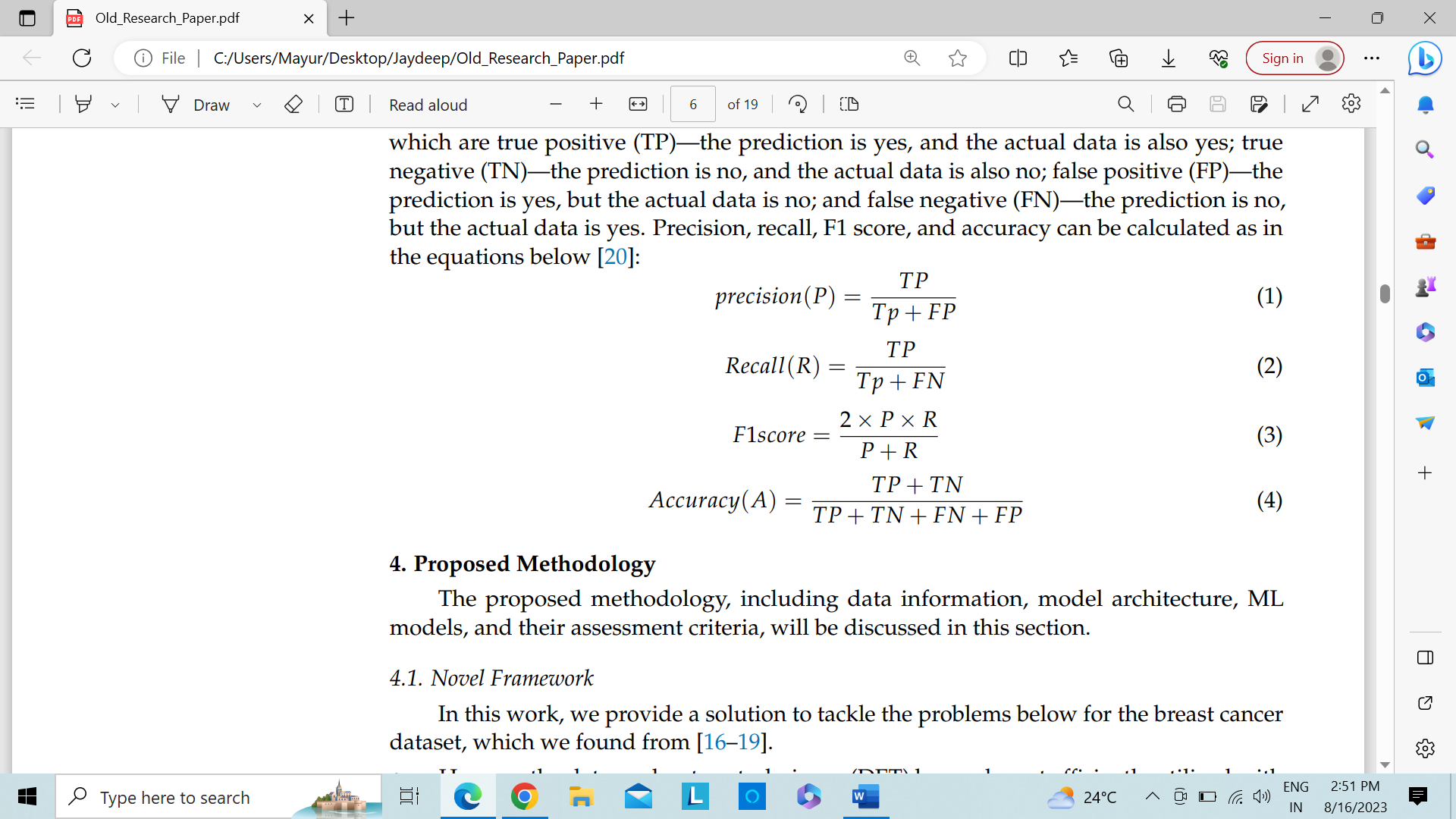


Table 1: WDBC dataset features classification

## 3.2 Performance Evaluation Matrices

In this study, we compared the precision, recall, F1 score, and accuracy cross-validation matrices. The numbers in the confusion matrix could be used to create these matrices since they remain TP, meaning that both the prediction and the observed data remain true; TN – both the prediction as well as the data support the negative conclusion; Both FP as well as FN situations in which the forecast is true but the actual data contradicts it. You may calculate P, R, F1-score, and A using the formulas below (L. Liu et al., 2023):



Where,

TP = true positive, FP = false positive, TN= true negative, FN = false negative

# 4. Proposed Methodology

## 4.1 Proposed Framework

We address the issues outlined under for the BC dataset that we discovered in this work (M. Zourhri; et al., 2023).

* How could DET be used in conjunction with BC prediction models to the greatest extent possible? (Rajaguru, H. 2019)
* How might BC features make ML models for cancer detection more accurate and scalable? (A. Rahman, Y. Zhang, J. G. Park, 2023) (D. Tsietso et al., 2023)

A solution is suggested in Figure 1 to address these issues. There are nine important steps in this solution. These are the broad strokes of this methodology:

1. Downloading the WDBC dataset from the ML repository.
2. Carry out the core data preparation operations.
3. Sort the data in WDBC into benign and malignant categories.
4. By determining their correlation, distribute the attributes into the three categories of positive, negative as well as random.
5. To get the best results, find less important characteristics and then remove such repetitive features.
6. Divide the dataset hooked on training as well as testing after EDA.
7. The deployment of 4-predictive machine learning algorithms on the WDBC dataset.
8. Following the execution of the models, the classifier's prediction is made using various matrices, such as confusion matrices, to gauge the efficiency of the models.
9. Review the findings and compare the precision of each model to earlier research.

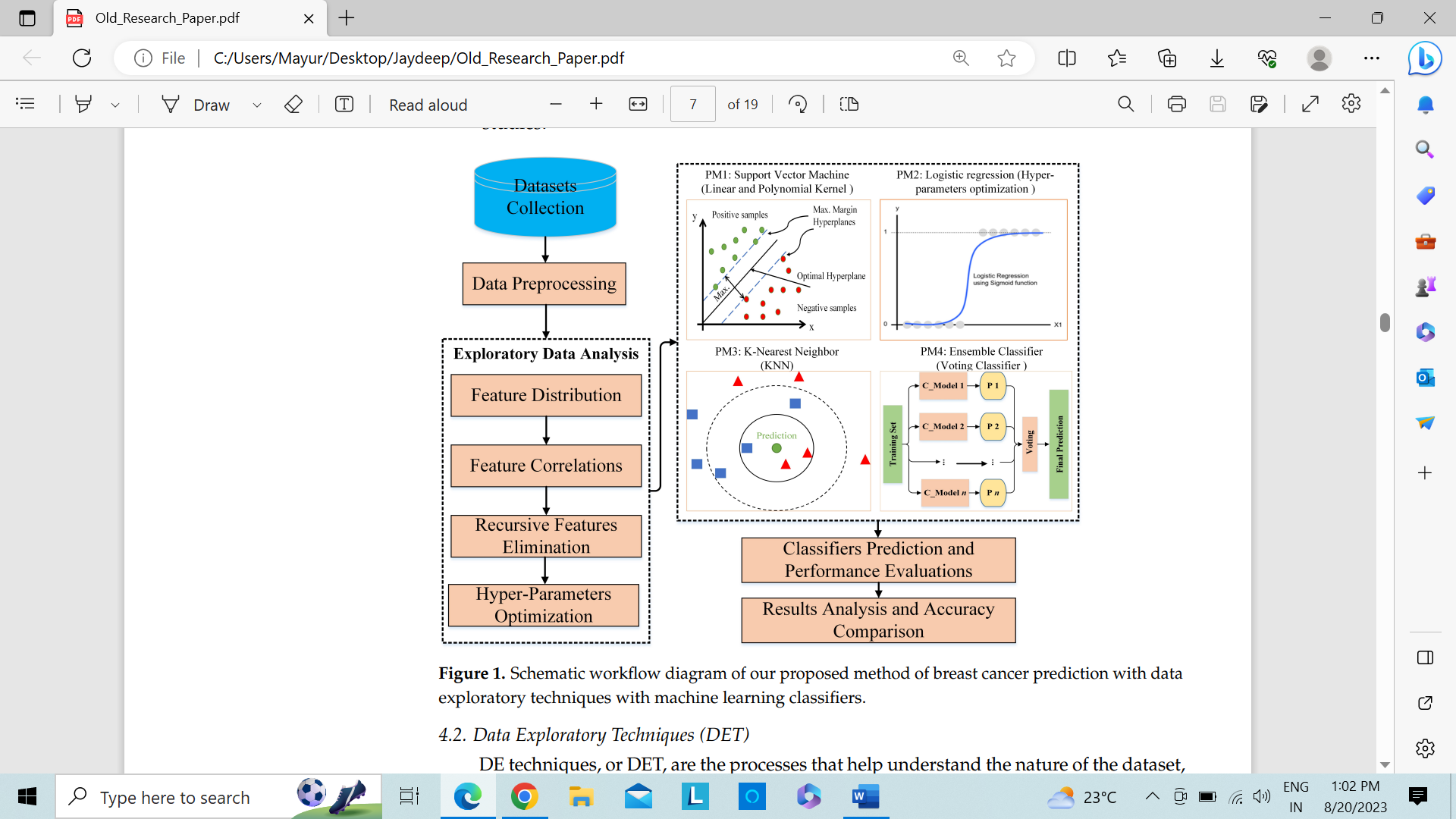


Fig 1.: Proposed diagram of BC using ML algorithms

## 4.2 DET (Data Exploratory Techniques)

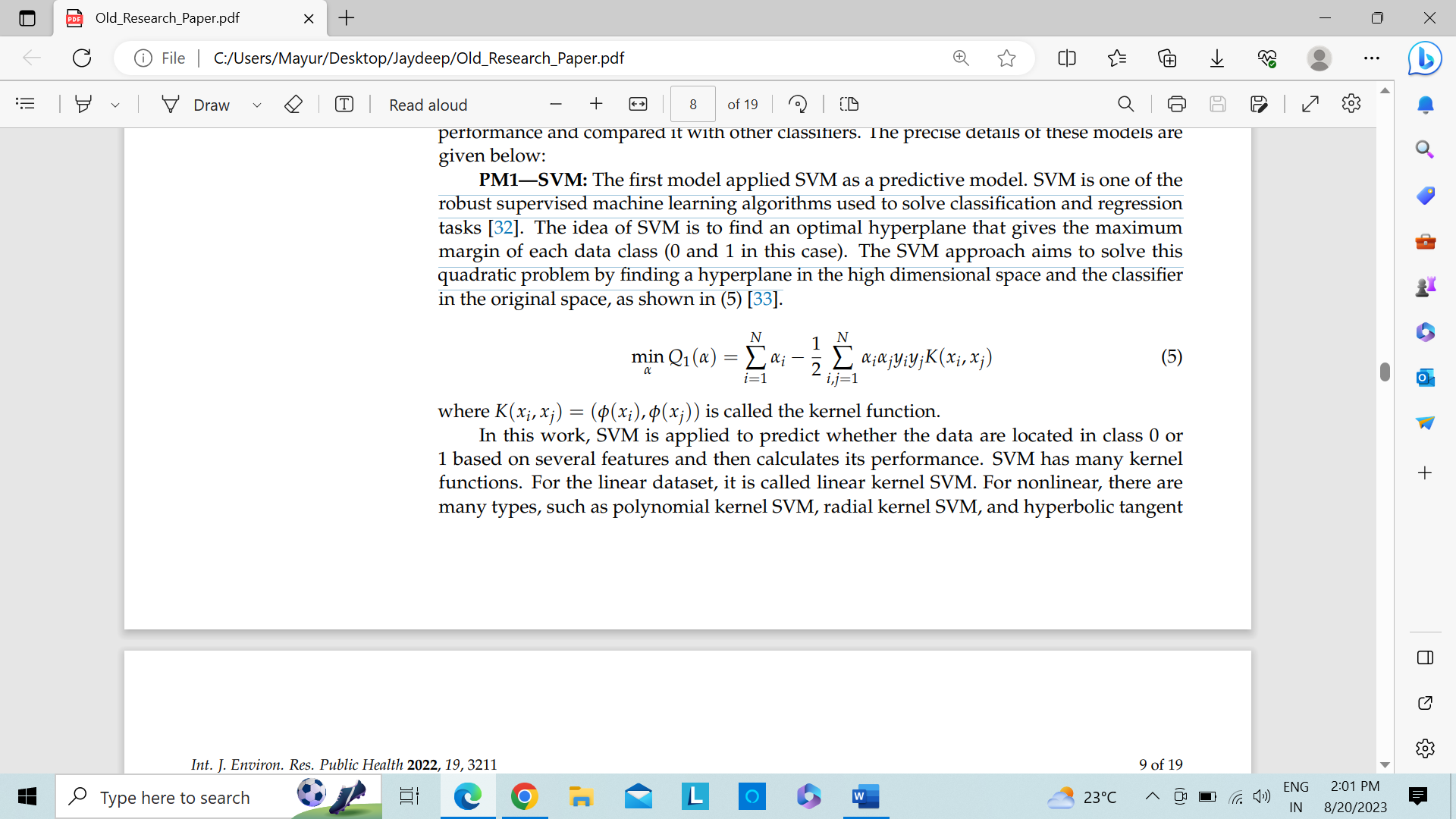
DET methods continue to be used to analyze the characteristics of the dataset and to find outliers or more easily accessible linked variables. Our study used recursive feature elimination, feature distribution as well as correlation coefficient as its DET.

1. **Feature Distribution:** In the WDBC dataset, the spreading of every feature selection remained first examined to determine how benign and malignant features differ from one another. Binary code was used to split the data into B = 0, M = 1, as well as A = 0, P = 1. Then, a plot between 0 and 1 was made showing distribution of every feature. Where, B = Benign, M = Malignant, A = Absence, P = Presence (Ramesh, S., *et al.,* 2022).
2. **Feature Correlation:** The CC between every of the 2-features is then calculated using the Pearson Correlation Coefficient (r) (Rasool, A.; et al., 2022), where, r = CC = Correlation Coefficient. Then, by classifying the features into 3-groups: 1). Positive 2). Negative 3). Uncorrelated — the relationship between the two features may be ascertained. If the variables are moving in the same direction, the feature selections would positively (r = +1) and negatively (r = -1) if these traits move in the other way (Sajiv G., Ramkumar G., 2022).
3. **RFE (Recursive Features Elimination):** One of the key procedures in ML is RFE. The dataset contains many features, thus choosing the quantity of feature selections that produce the best possible prediction results is crucial for enhancing model performance. The main goal of RFE is to use fewer elements while yet providing better understanding (Jabeen, K.; et al., 2022). The loop will continue until the ideal number of features is discovered. It was carried out by the SkLearn function selector. Fit (x, y). Using all the feature selections in training dataset as a starting point, RFE attempts to find a subset of features by successfully deleting features one at a time until the target number of features is left. This is accomplished by re-fitting the model using the supplied ML algorithm, ranking the features according to their importance, and removing the least important features. Up until specific amounts of features are kept, this process is repeated (Ebrahim, M.; Sedky, A.A.H.; Mesbah, S., 2023).
4. **HO (Hyper Optimization):** HO is a machine learning technique used to fine-tune a set of ideal parameters. The learning process is managed by these parameter values. There are numerous methods for optimizing hyperparameters, including GS, RS, BO, gradient-based optimization, and population- and evolutionary-based methods (Wang, X.; et al., 2022). Due of grid search optimization's successful outcomes for optimization, we applied it in this investigation. From a random\_search\_grid of boundary values provided by the parameter; candidates are generated using the brute-force method. Obtaining the greatest cross-validation confusion metric values is the aim of the grid search. Due to the WDBC prediction datasets, we used a GridSearch K-fold CV based on SkLearn in this instance. GridSearch CV was used to assess the hyperparameters in all prediction models (Hanan Aljuaid, et al., 2022).

## 4.3 Machine Learning Predictive Algorithms

Four different ML classifiers were used in this study as predictive algorithms to determine if the Y-variable in the data in the WDBC dataset was malignant or benign. Data were split into training and testing sets, each containing 80% of the data. In our trials, we used an integer number as the random\_state to generate this distribution (Abunasser, et al., 2022). This variable can be any value to tweak the hyperparameter, however split\_size must be a specific number. The models were built using the training WDBC dataset, and their performance was then assessed using a test dataset. We choose SVM since it had the greatest accurateness in the prior research, and by adjusting the hyperparameter, LR performed at its best. KNN was also chosen since it produced good results with the input features (Ortega-Martorell, S., Riley, P., Olier, I. *et al.*, 2022). As Using the ensemble-based classifier as an implemented, we evaluated its performance using voting approaches, and we compared it to other classifiers.

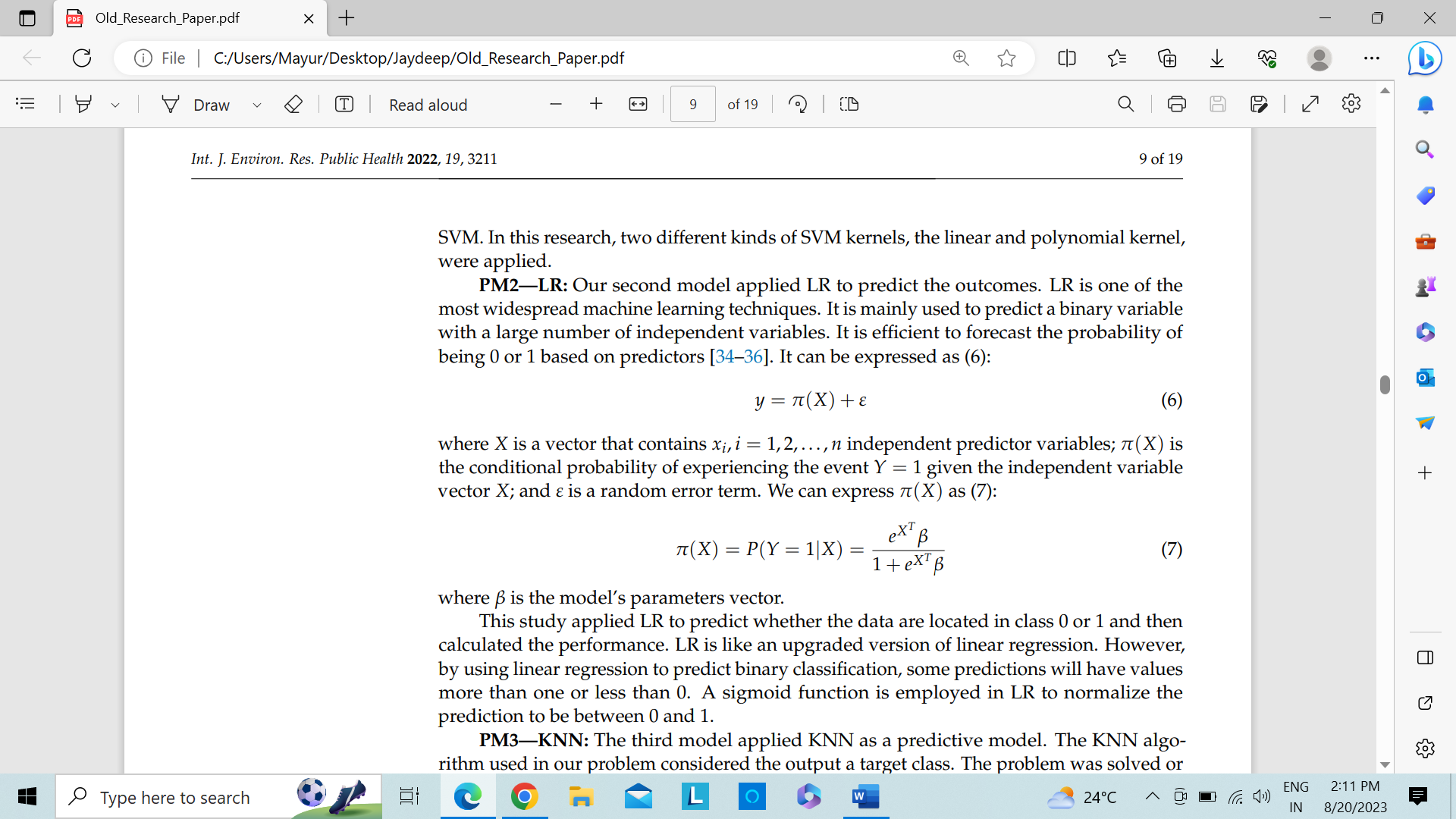
1. **Support Vector Machine (SVM):** SVM was used as a predictive algorithm in the first model. One of the reliable supervised machine learning methods for classification and regression is SVM (D. Rajasekhar; et al., 2023). Finding an ideal hyperplane that provides the largest margin for every data class is the goal of SVM. By locating a hyperplane in the high-dimensional space as well as revealed in below formula (S. Goudarzi, et al., 2023), the SVM technique seeks to overcome this quadratic problem.



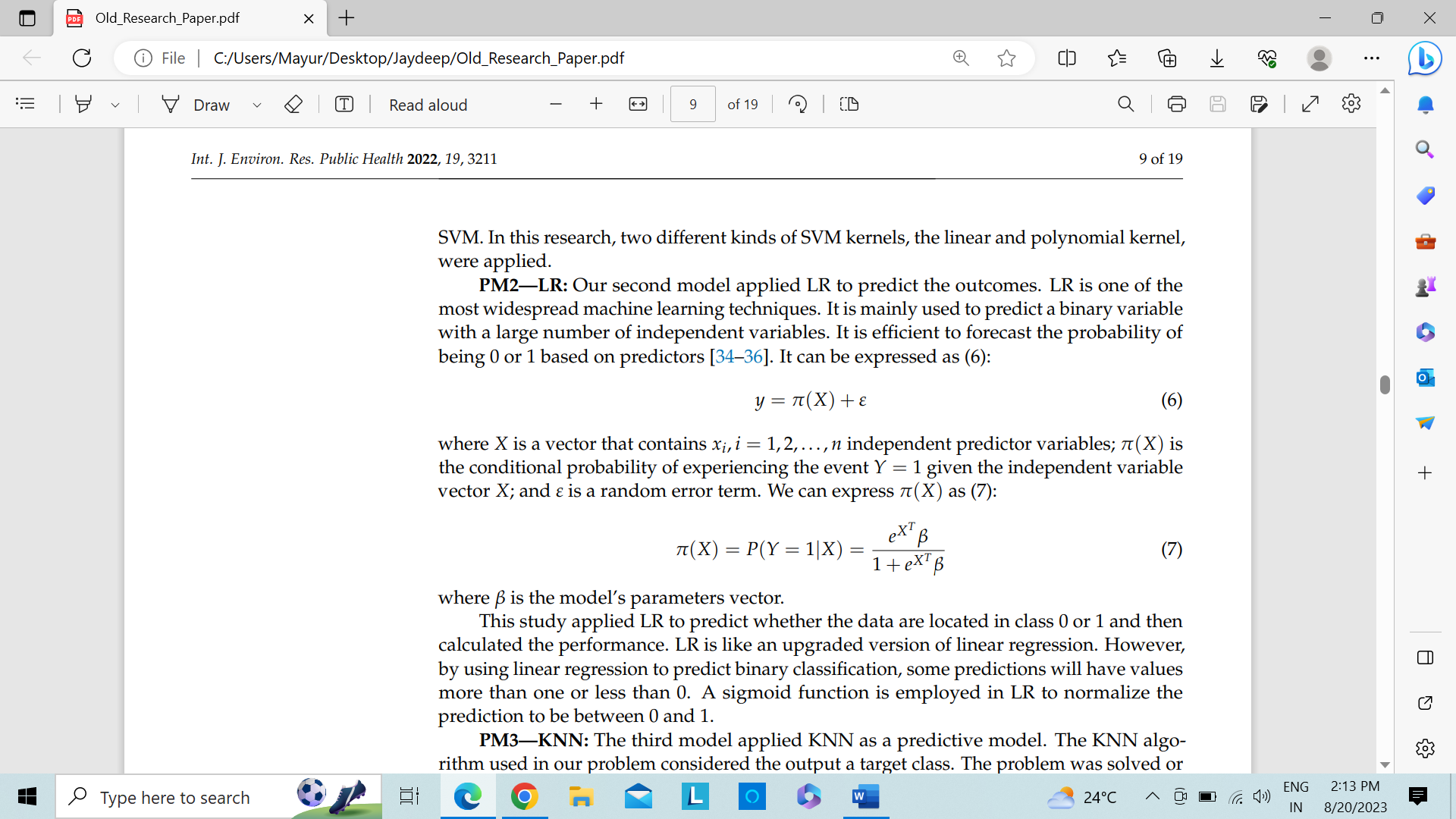
The above formula is called the kernel function.

SVM is used in this work to forecast whether the data are in binary class based on several features, and its performance is then calculated. SVM's kernel features are numerous. It is known as linear kernel SVM for the linear dataset.

**2. Linear Regression (LR):** The outcomes were predicted using LR in our second model. One of the most popular machine learning methods is LR (Al-Azzam, N.; Shatnawi, I., 2021). Prediction of a 0 and 1 value with many independent factors is its principal use (A. U. Haq et al., (2022). Forecasting the likelihood of being binary variable based on predictors is effective (Hasan, A.S.M.T.; et al., 2022). This can be written as (6):



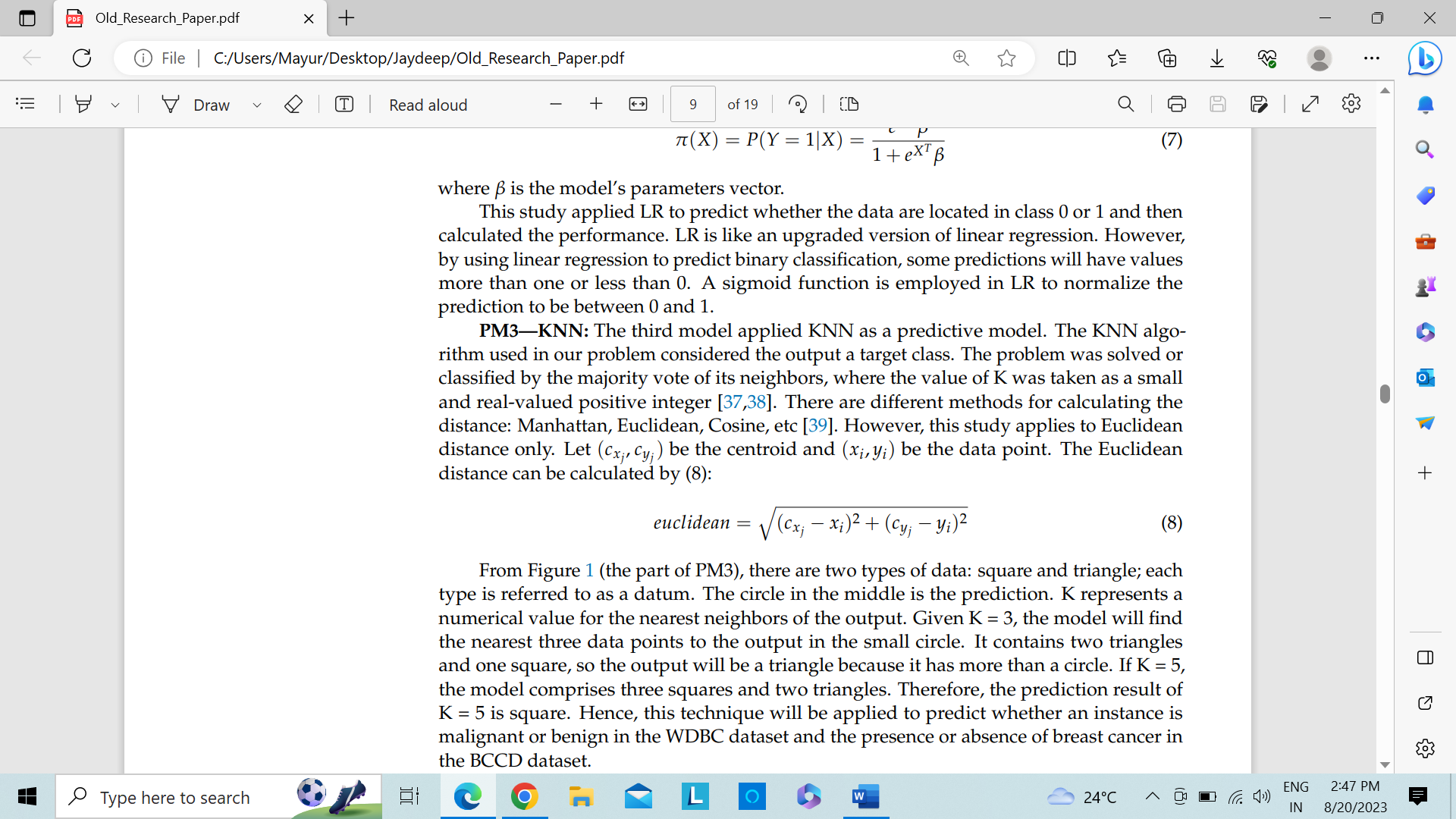
We could write (7) for π (X):



Where, β = parameter vector of the model

The performance was then calculated after using LR to forecast whether the data was in class 0 or 1. LR resembles a more advanced kind of LR. However, certain predictions made using LR to forecast binary classification will have values greater than one or lower than zero. To normalize the prediction values between 0 and 1, LR uses a sigmoid function.

**3. K-Nearest Neighbor (KNN):** The output was treated as a target class by the KNN algorithm employed to solve our issue. By a mainstream vote of the situation neighbors, the issue was resolved or categorized (Mejdoub, M.; Amar, C.B., 2013) (Yu, Z.; et al., 2015). K was assumed to be a tiny, positive integer with a real value. The distance can be calculated using a variety of techniques, including Manhattan, Euclidean, Cosine, etc. (Soham Chattopadhyay; et al., 2022). The study solely uses Euclidean distance, though. Let the centroid be (cxj, cyj) and the data point be (xi, yi). (8) can be used to calculate the Euclidean distance.



Two different forms of data—a square and a triangle—from Figure 1 (the portion of PM3) are referred to as datums. The forecast is shown in the round in the center. K = output's nearest neighbors. The model will locate the three data points in the tiny circle that are closest to the output, given that K = 3. It has more than a circle because it has 2-triangles as well as 1-square, hence the result will be a triangle. The model has 3-squares as well as 2-triangles if K = 5. As a result, K = 5's forecast result is square. Therefore, this method will be used to determine if an occurrence in the dataset of WDBC is malignant or benign.

**4. Ensemble Classifier (EC):** To find every malignant tumor in the WDBC dataset, it seeks to maximize precision and recall. In our study, an ensemble classifier was used to enhance the LR algorithm (Hager Saleh, et al., 2022) (M. Liu et al., 2022). EC come in a variety of forms, including voting, boosting, and bagging (S. Gupta; et al., 2022). The voting classifier is the one that will be applied in this study. Then, we applied them to the similar dataset to obtain each model's prediction outcome. To decide a last prediction, a majority vote is required. As an illustration, the EC trained three different algorithms: algo.1 -> 1, algo.2 -> 0, and algo.3 -> 0. Since two of them are "0" and only one has a different outcome, the result would be "0".

## 4.4. Experimental Setup

Python language was used to carry out this task. We carried out the important procedures that the data analyst or doctor can use to apply this effort for the real-time BC prediction:

1. Run the preprocessing procedures to remove the missing values after importing the relevant Python libraries.
2. Process and apply each dataset's four-layered data exploration approaches.
3. Definitions and identification of each relevant functions, including cross-validation metrics, the confusion matrix, the precision-recall curve, the ROC curve, and the learning curve.
4. Put the proposed architecture prediction algorithms into practice:
   * Starting with SVM, it is necessary to provide the variables and the quantity of training and test sets. The model should next be performed utilizing such as Linear SVM as well as Polynomial after defining the output findings. Cross-validation metrics would be used to display the results.
   * The LR model is the one that comes next. After specifying the variables and dividing the data, two approaches were used to determine the ideal hyperparameter. The first one included using GridSearchCV, while the second involved RFE. Then, determine which crossvalidation metrics were used to both approaches by plotting the ROC curve, the confusion matrix, and learning.
   * The third prediction model was KNN, and to discover the optimal hyperparameter for KNN, we utilized GridSearchCV. We also displayed the confusion matrix, as well as cross-validation metrics.
   * The final model is EC; for this work, LR was combined with EC as well as VC. The stages involved in execution are like those from earlier. The CM, LC as well as C-V metrics display the results.

Table 2 presents the experimental setting and key software packages for putting the offered prediction algorithms as well as DE approaches into practice.

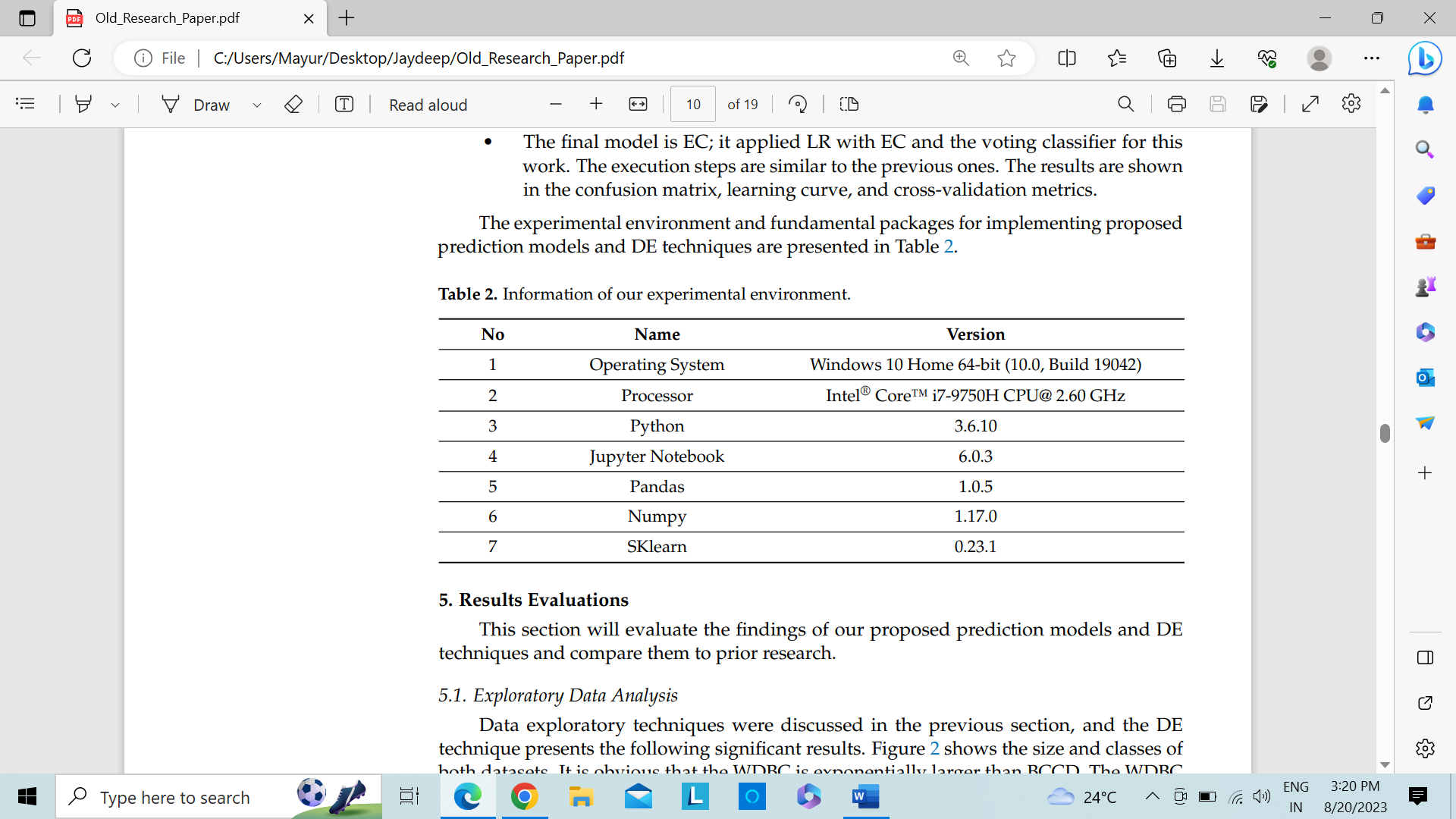


Table 2 – Testbed of Implementations

# Results and Discussions

This section will assess the results of our suggested DET and prediction models and compare them to earlier studies.

## 5.1 EDA:

DET were covered in the part above, and the DE method yields the noteworthy findings listed below. The two dataset sizes and classifications are depicted in Figure 2. There are benign as well as malignant modules in the WDBC. Therefore, a thorough investigation of the WDBC will yield more knowledge. To illustrate the dataset, this study specifically concentrated on averages, SE, worst as well as correlations.

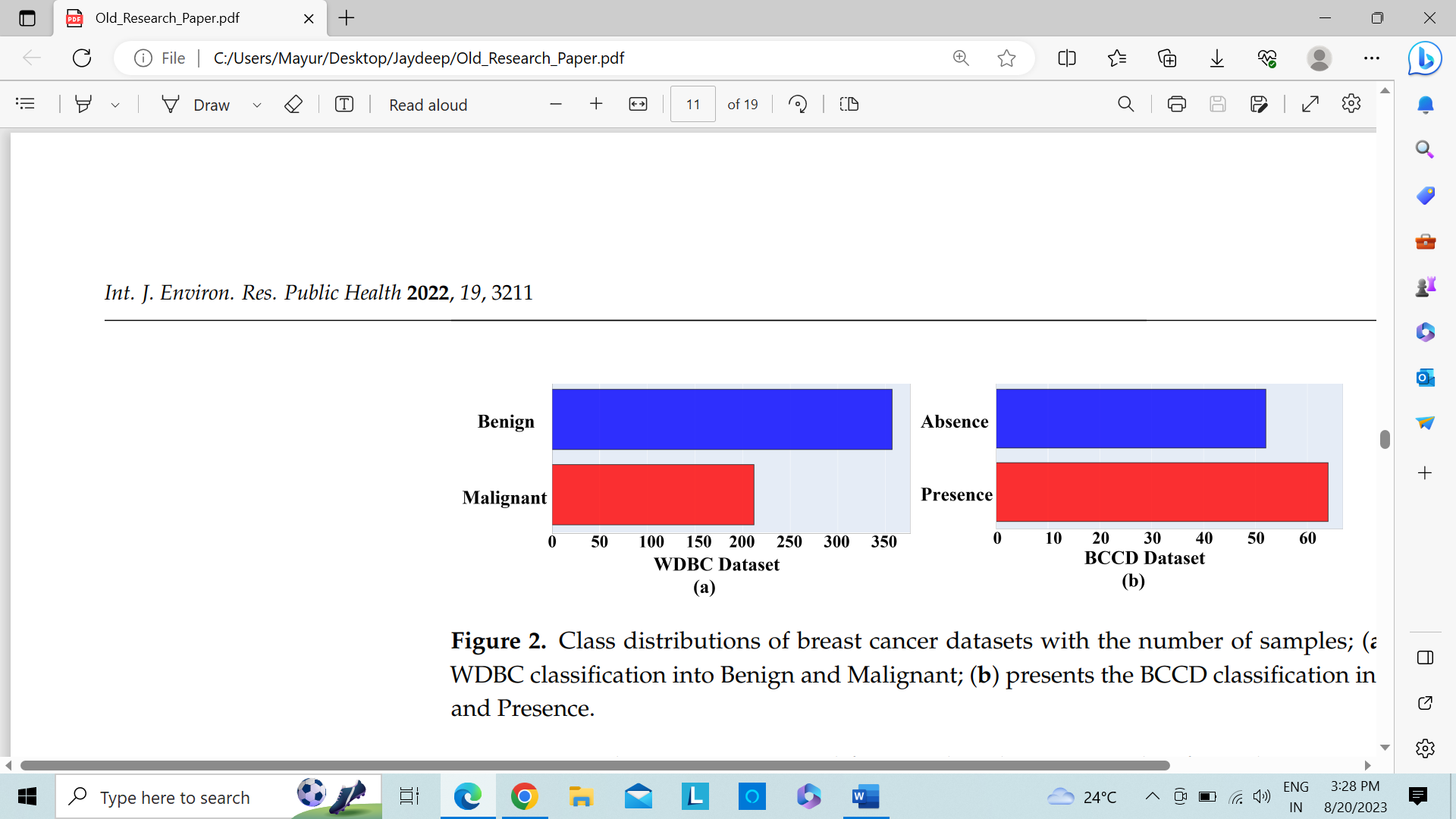


Fig 2.: The BC dataset's class distributions show that WDBC is classified as benign and malignant

In figure 3, we chose two samples at random from each feature. In figure 3(a) -> radius means -> the benign and malignant WDBC classes are shaped differently, with the former showing the highest intensity. The concentration level in together classes remained nearly comparable in form in texture mean (b), but otherwise. It also clarifies the SE Smoothness analysis of feature selection dataset based on smoothness and concave points. Both the benign and malignant concave up and down graphs (c) were different, and the inflection point crossed the peaks and valleys. Concave down and up, however, were roughly in the same ranges in SE smoothness (d), although malignant slopes were greater than benign slopes. Based on texture and area, it displays the worst feature such as (e) and (f) distribution. Here, benign and malignant texture waves have a similar appearance. Malignant cells are, once more, flatter and longer in the area graph.

We also provided the remaining curves for the feature in Note 01 of the Supplementary Materials.

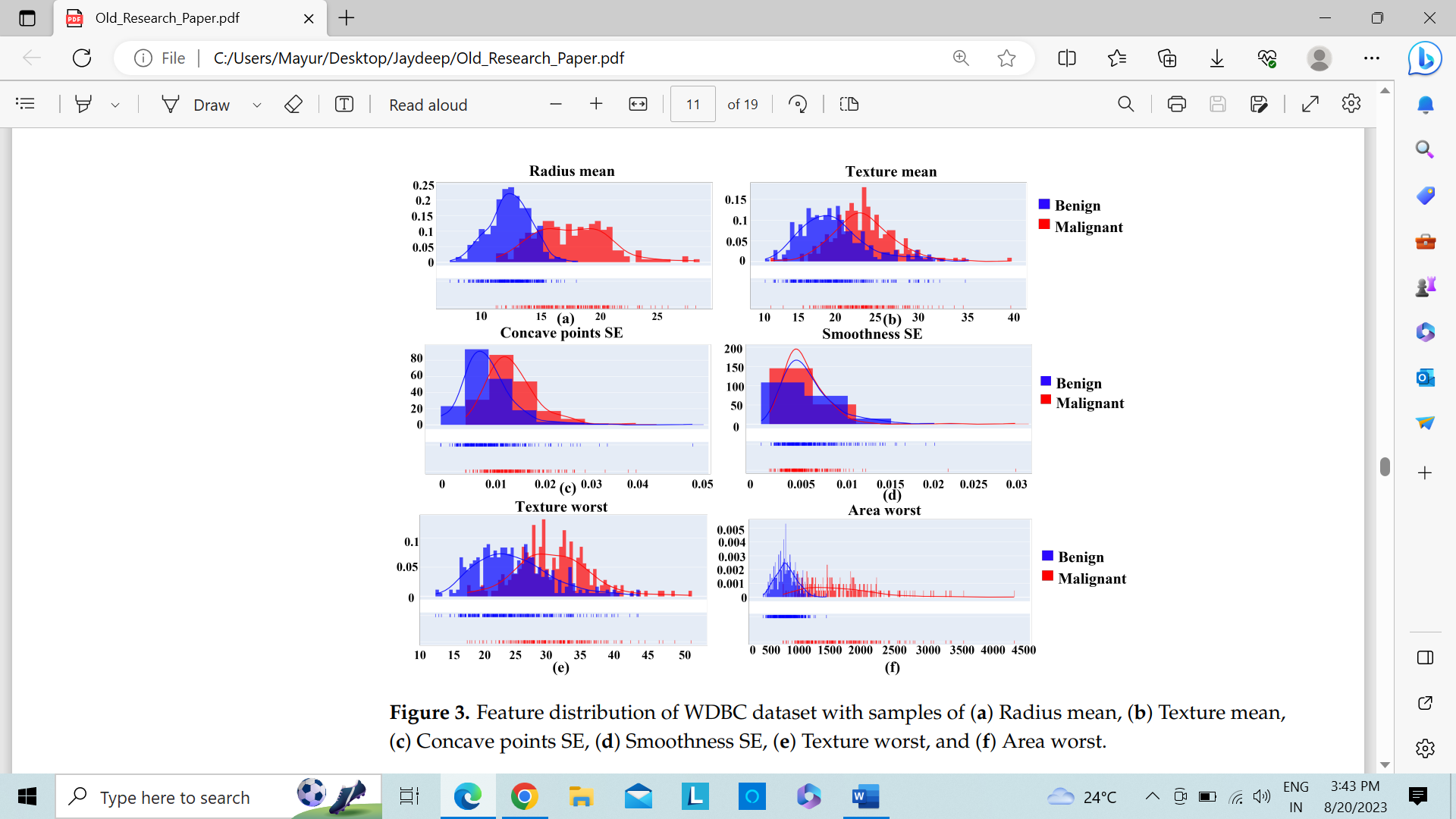


Figure 3: Distribution of features in the WDBC dataset using various samples

Additionally, Figure 4 displays the feature selection relationships created on characteristics that are favourably 4(a), 4(b) and 4(c) across various features and samples. For instance, there is no meaningful link between Texture worst as well as Symmetry means (b). Due to the limited amount of room in this study, we only showed a few feature matrices. In conclusion, the proposed prediction models were able to more accurately detect the tumour thanks to the correlational feature distribution 80% of the data analysis. It is important to note that training took up 80% of the overall data set, and test data acquired up 20% of the data set. Utilising a colour as well as value scheme to discriminate between positively, uncorrelated, and negatively correlated features, it displays the correlation of each pair of features. For instance, the WDBC dataset shows that area mean and radius mean have a positive correlation, texture mean as well as smoothness mean have a negative correlation as well as smoothness SE as well as radius mean have a negative correlation.

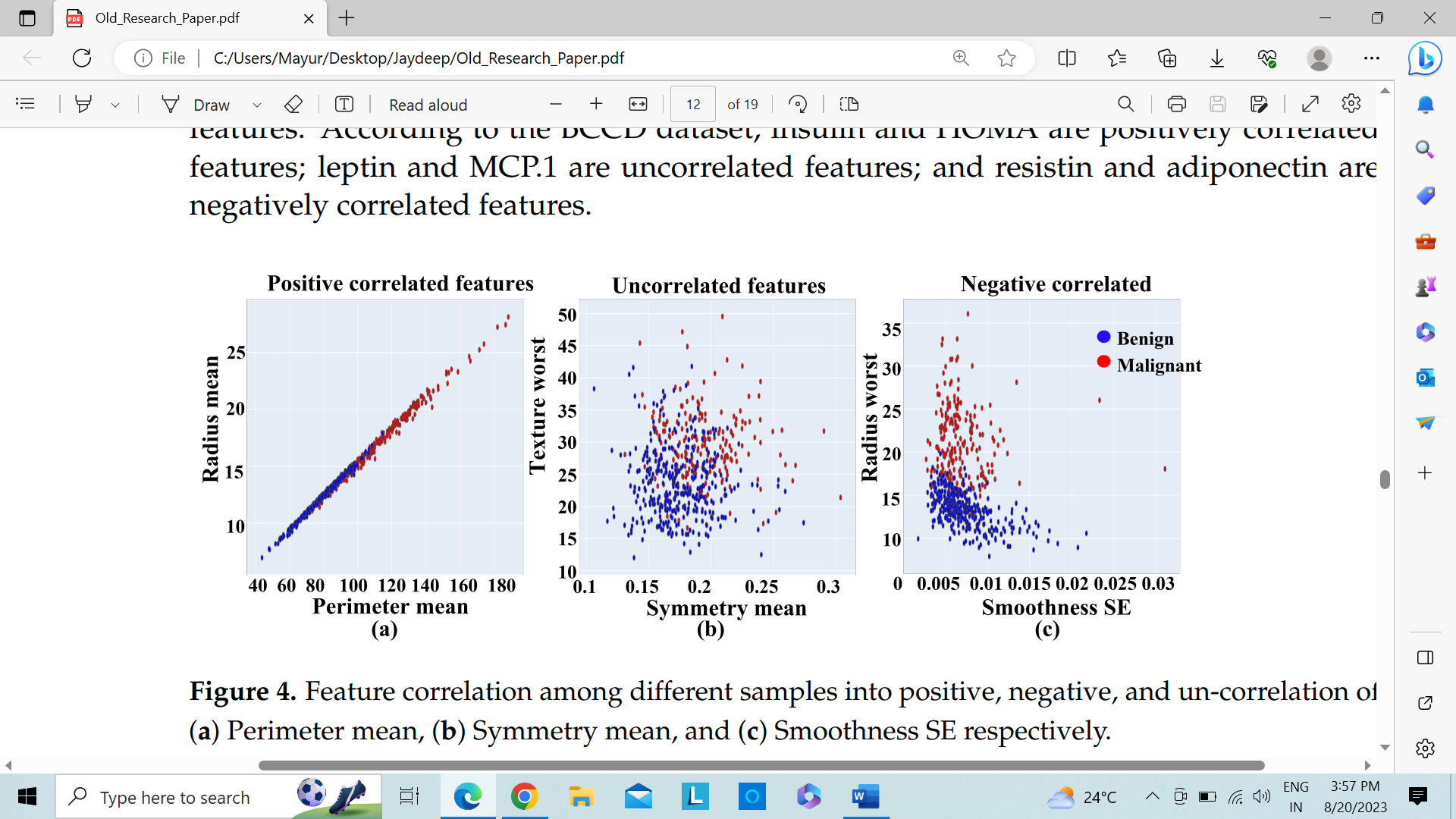


Figure 4.: Feature correlation between several data samples such as Perimeter, Symmetry as well as Smoothness

## 5.2 Machine Learning Predictive Algorithms Evaluation

### 5.2.1. Support Vectoe Machine (SVM):

The linear kernels as well as polynomial kernels were two SVM utilising SVMs that were taken into consideration in this work. Bolder entries in Table 3 represent the highest achievements for the execution analysis of both linear kernels as well as polynomial kernels through confusion matrices. The polynomial kernel performed better on the WDBC data set than the linear kernel on both the training and testing sets. The polynomial kernel's precision score in training sets was practically identical to that of the linear kernel, while its F1 score and accuracy score were both 99.12%. Figure 5 shows a performance comparison of the two kernels using four cross-validation scores using the WDBC dataset, with the performance analysis of the polynomial kernel being shown to be superior.

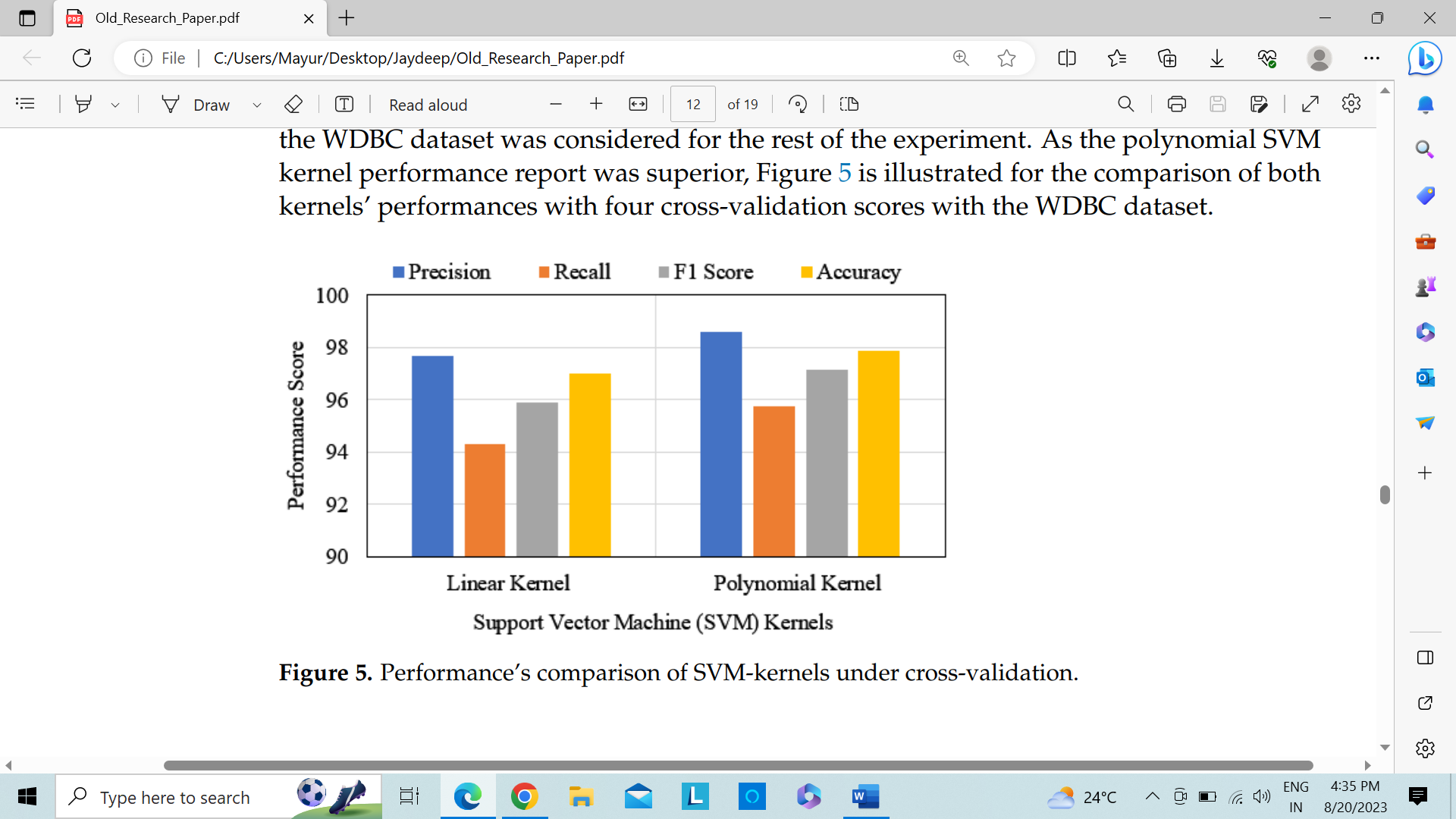


Figure 5: SVM-kernels' performance is compared via cross-validation

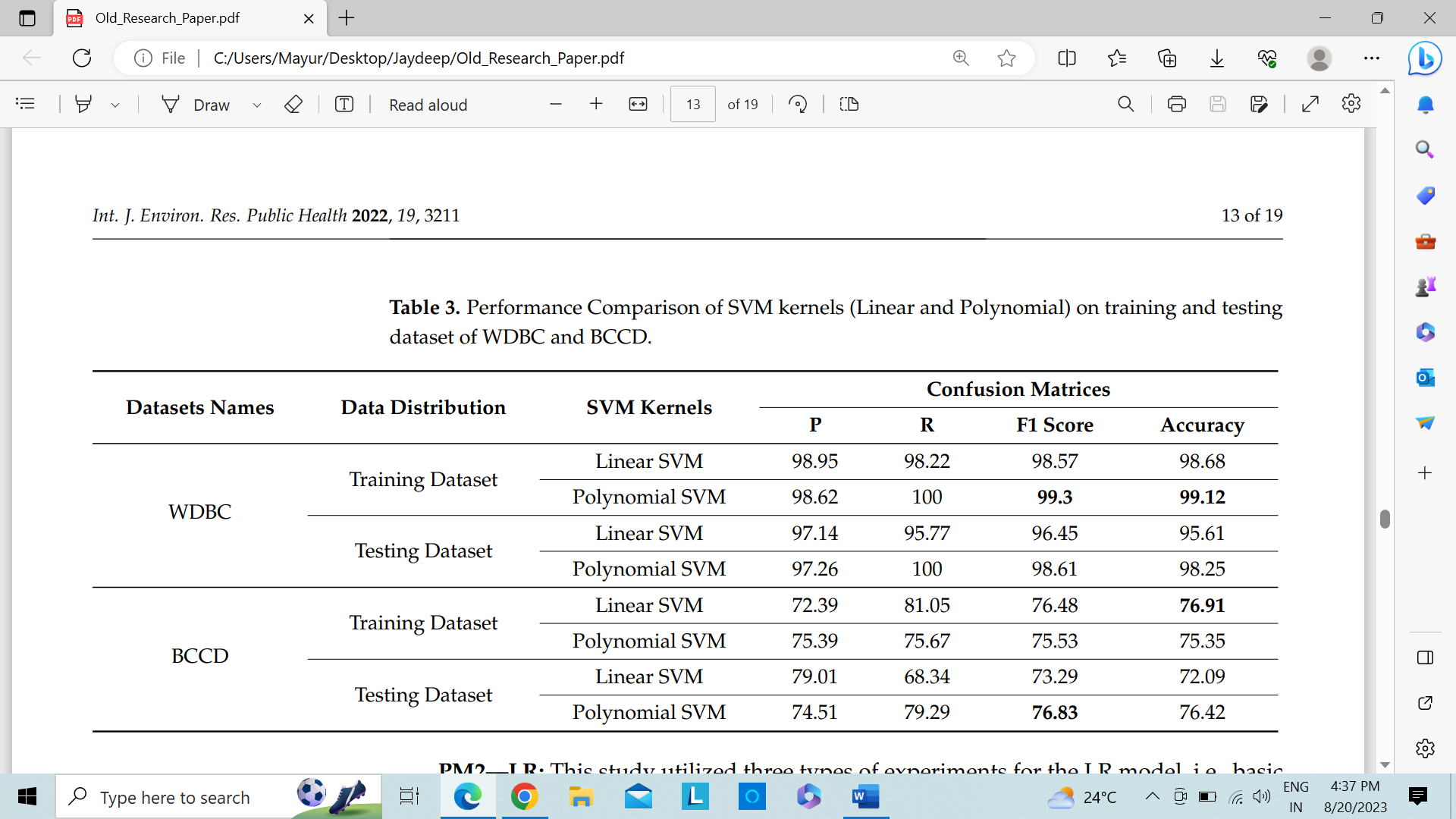


Table 3: Comparison of the performance of linear and polynomial SVM kernels using the WDBC training and testing datasets

### 5.2.2. Linear Regression (LR):

On the WDBC dataset, this study used three different types of LR experiments: basic LR, LR-recall as well as LR-RFE approach. Figure 6 contrasts the LR learning curve's performance with RFE approach, though. The training scores for both models were significantly far more important than the C-V scores for the limited amount of data. The generalisation of the cross-validation score and training score will, however, most certainly grow with the addition of more training samples. With the larger sample size, LR with the RFE model (b) produced better training and cross-validation results. Furthermore, the scores were approaching each other more closely than in the basic LR model (a). The C-V execution analysis is shown in Table 4, where bolder entries represent the highest achievements. The performance of LR with RFE was the best of these three approaches, with accuracy and F1 score results of 97.36% and 98.06%, respectively. The simple LR performed second best with a little lower matrix score. Conversely, LR with a 100% recall rate received the lowest ratings.

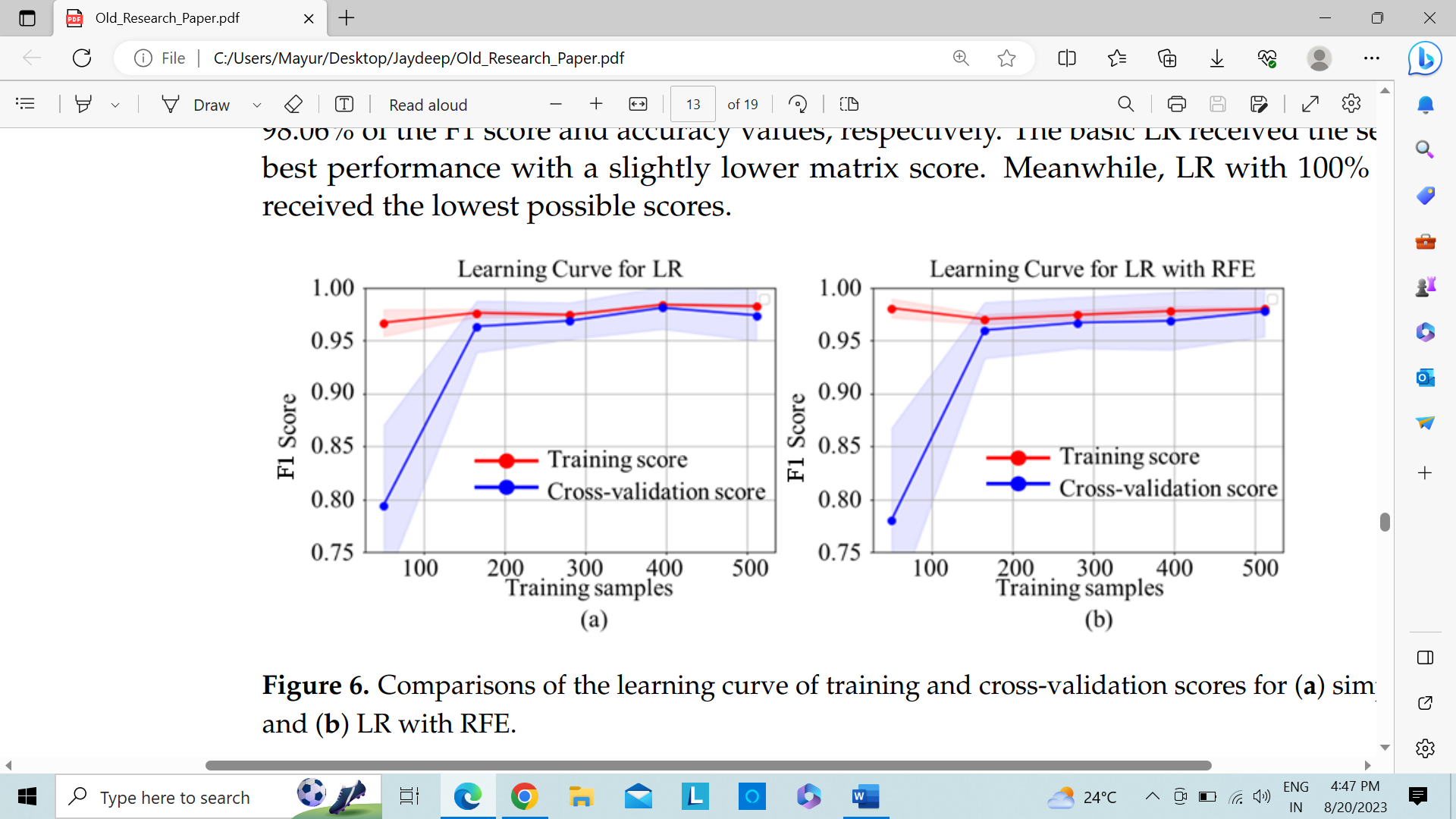


Figure 6: Comparisons of the cross-validation scores and learning curves for (a) LR (b) LR with RFE.

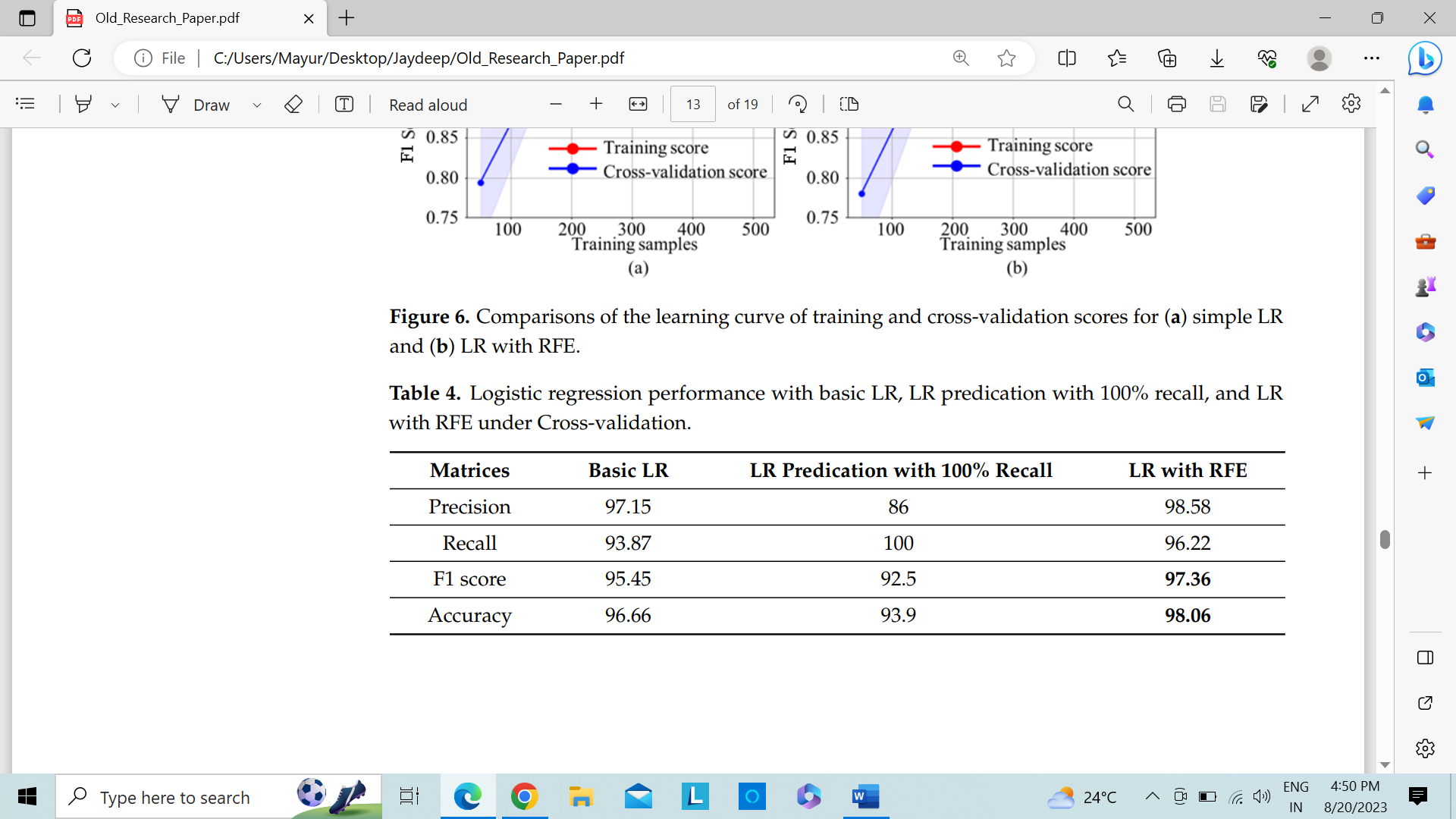


Table 4: LR, LR-recall predication, LR-RFE as well as performance with basic LR-cross-validation

### 5.2.3 K-Nearest Neighbor (KNN):

KNN with hyperparameter and basic KNN have both been tested by the KNN predictive model. Figure 7 makes it obvious that KNN with hyperparameter performed better than basic KNN. Basic KNN automatically runs with default settings and outputs results. Hyperparameter, however, enables KNN parameter tweaking. It shows that the fundamental KNN obtained a 95.43% accuracy and an F1 score of 94.73%. KNN with hyperparameter, meanwhile, scored 97.35% F1 and had 97.01% accuracy.

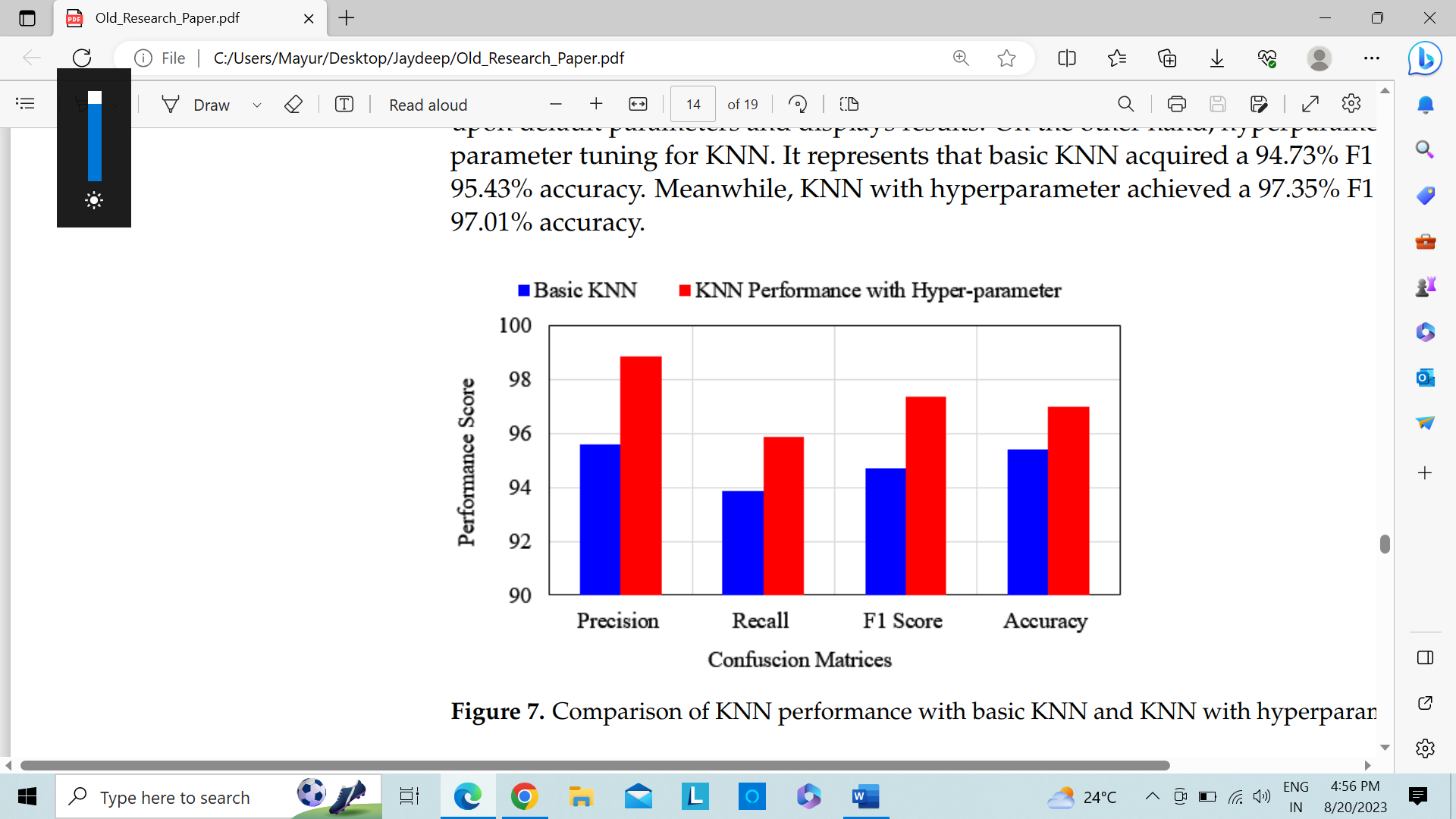


Figure 7: Performance comparison between KNN with hyperparameter and KNN with basic KNN

### 5.2.4 Ensemble Classifiers (EC):

Table 5 presents the execution analysis of EC, where bolded entries represent the best results. The WDBC dataset is evaluated using three different approaches: voting classifier, E-LR as well as prediction of Voting Classifier-recall. When compared to CV prediction with 100% recall, the ensemble LR and CV produced the best results. With the provided dataset, it is obvious that CV was effective in achieving a F1 score is 96.02% as well as 97.61% accuracy. Similarly, group LR performance is important as well. On the other hand, a 100% recall results for the CV did not produce any useful results.

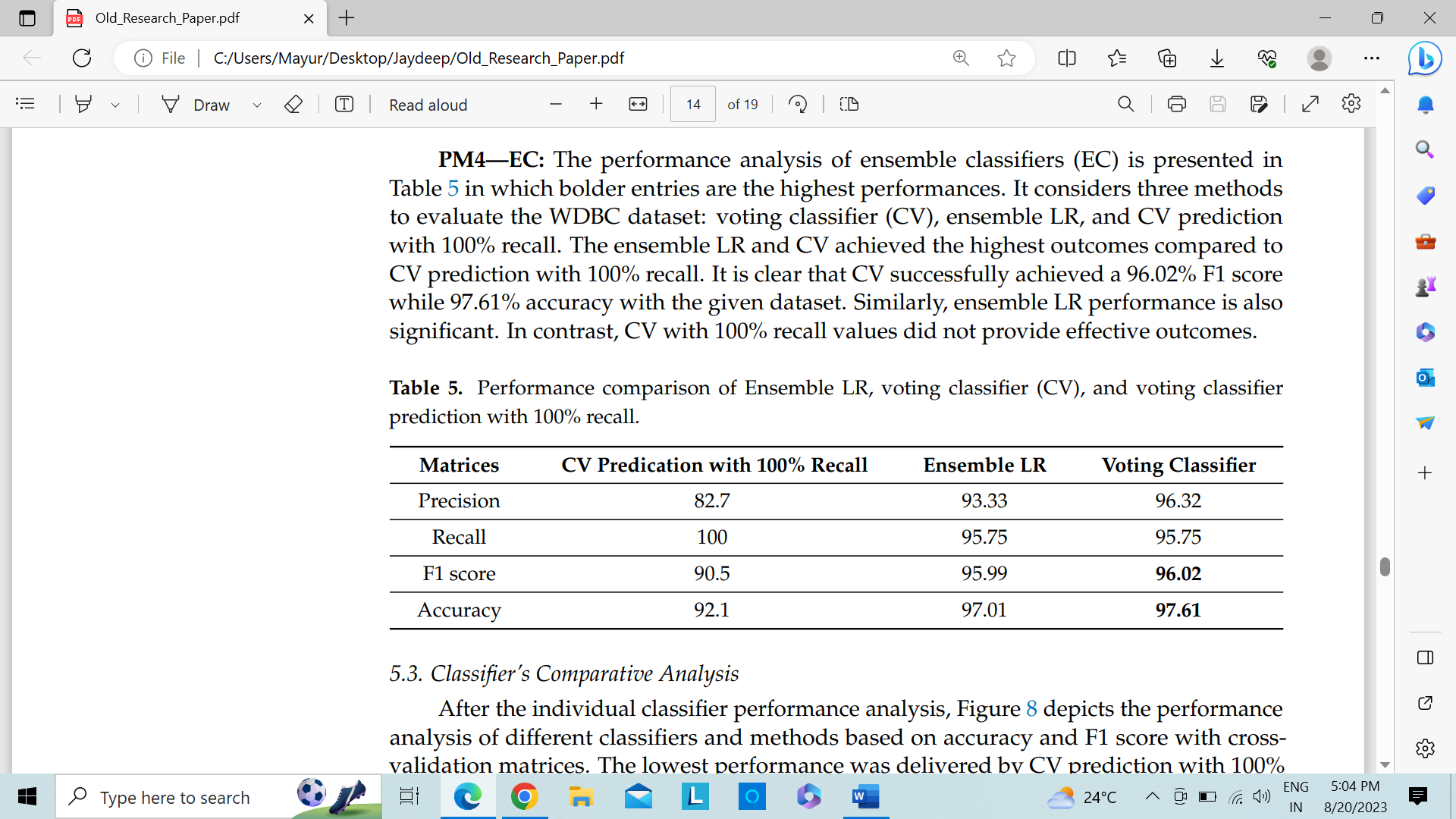


Table 5: Voting classifier-recall, E-LR, and voting classifier (CV) performance comparison

## 5.3 Analysis of Classifier's Comparability

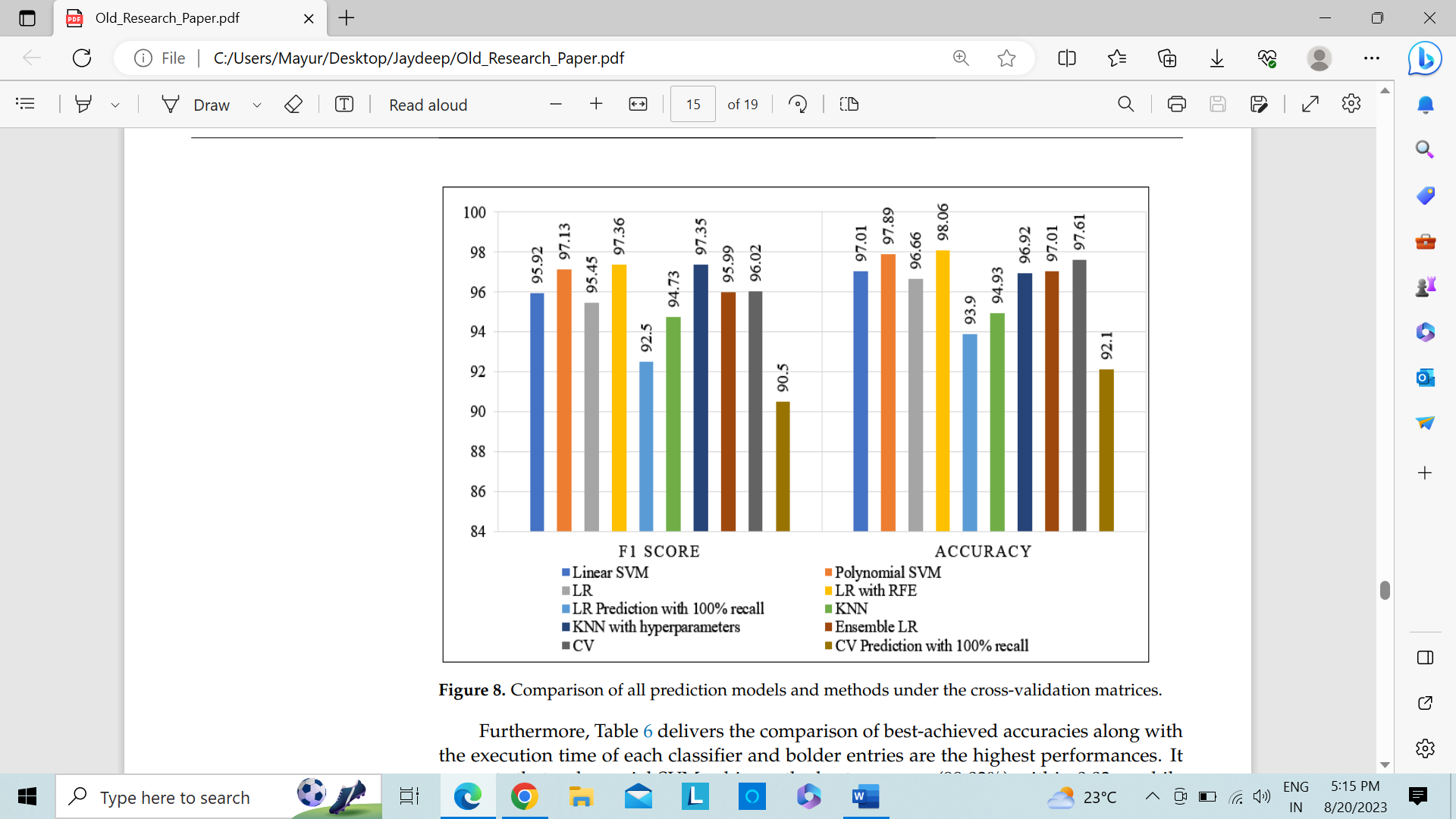
Figure 8 shows the performance analysis of various classifiers as well as methods based on accuracy as well as F1 score through crossvalidation matrices after the analysis of the performance of each individual classifier. With an F1 score and accuracy values below 95%, LR, K-Nearest Neighbor as well as CV-Recall (100%), prediction with 100% recall had the worst performance. LR with RFE outscored other approaches in this comparison, achieving the accuracy was 98.06% as well as F1 score was 97.36%. The performance of Support Vector Machine polynomial, Voting Classifier as well as K-Nearest Neighbor through hyperparameters is advantageous. Therefore, it is evident from these analyses that RFE with LR performs better in cross-validation than all other approaches combined.

Figure 8: All prediction models and techniques are compared using cross-validation matrices

Additionally, Table 6 compares the top accuracies obtained as well as the execution times of each classifier, with bolded entries representing the highest performances. It shows that basic KNN performed in the quickest time but with the lowest accuracy, whereas polynomial SVM performed with the greatest accuracy was 99.03% by the time was 0.03 s. However, LR using the RFE approach performed admirably and achieved the greatest cross-validation accuracy (98.06%). However, while having a higher accuracy (97.35%), KNN with hyperparameter had the longest execution time (4.023 s).

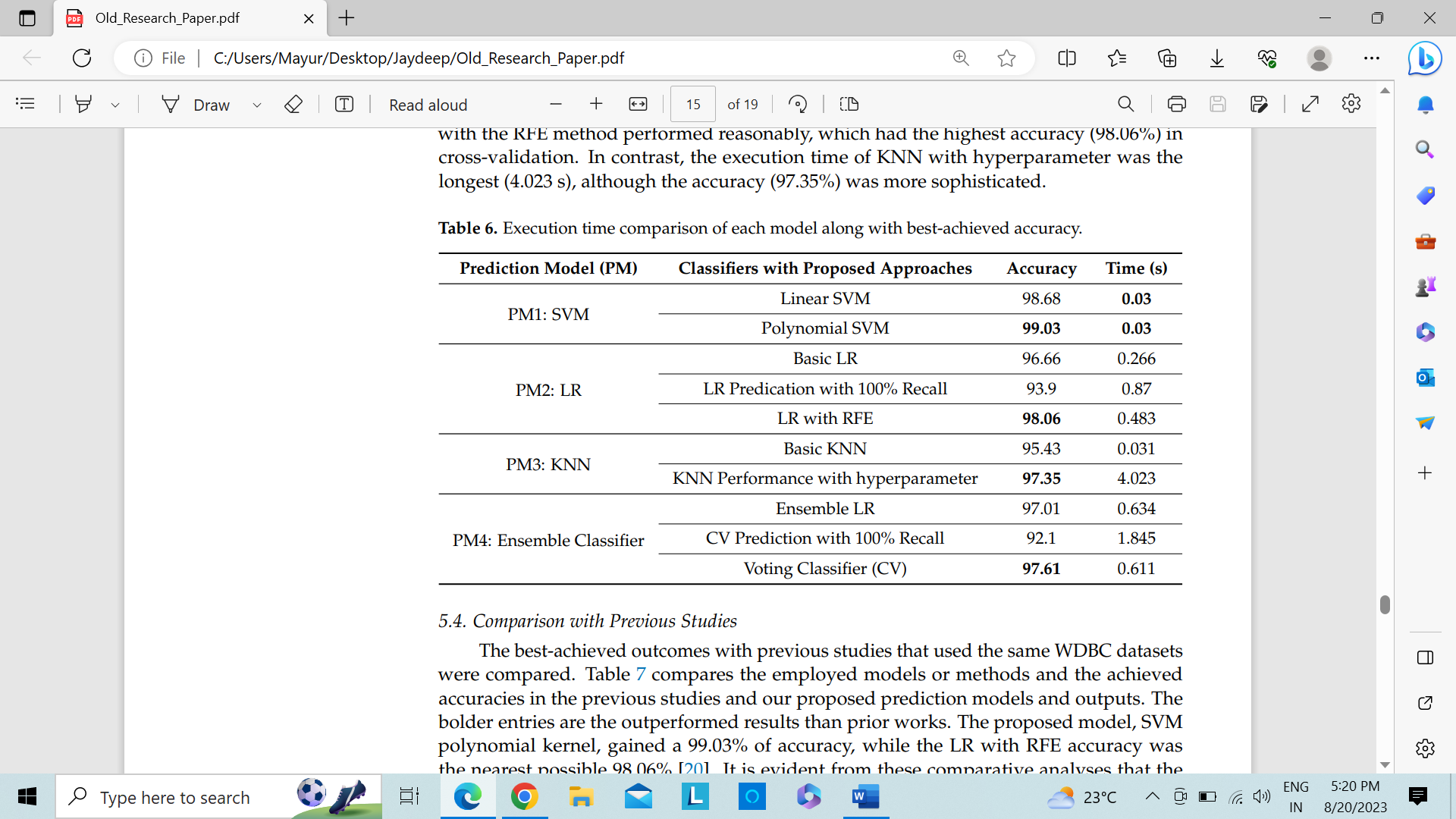


Table 6: Comparison of the execution times for each model and the highest level of accuracy

## 5.4 Compared to earlier studies

Comparisons were made between the best results from earlier studies that made use of the identical WDBC datasets. Table 7 contrasts our suggested prediction algorithms as well as and outputs with those used in the prior investigations, as well as the accuracy levels attained in those studies. The outcomes that exceeded earlier efforts are represented by the bolder entries. The accuracy of the suggested model, the SVM polynomial kernel, was 99.03 percent, whereas the The accuracy of LR-RFE was as close as it could go, 98.06% (Brause, R.W., 2001). These comparison evaluations clearly show that the suggested prediction models beat the earlier methods and attained a level of accuracy suitable in order to find BC. The proposed DM approaches with the ML prediction models may be the cause of remarkable advancements in comparison to earlier studies. The DE approaches allowed for the greatest precision while using the least amount of execution time.

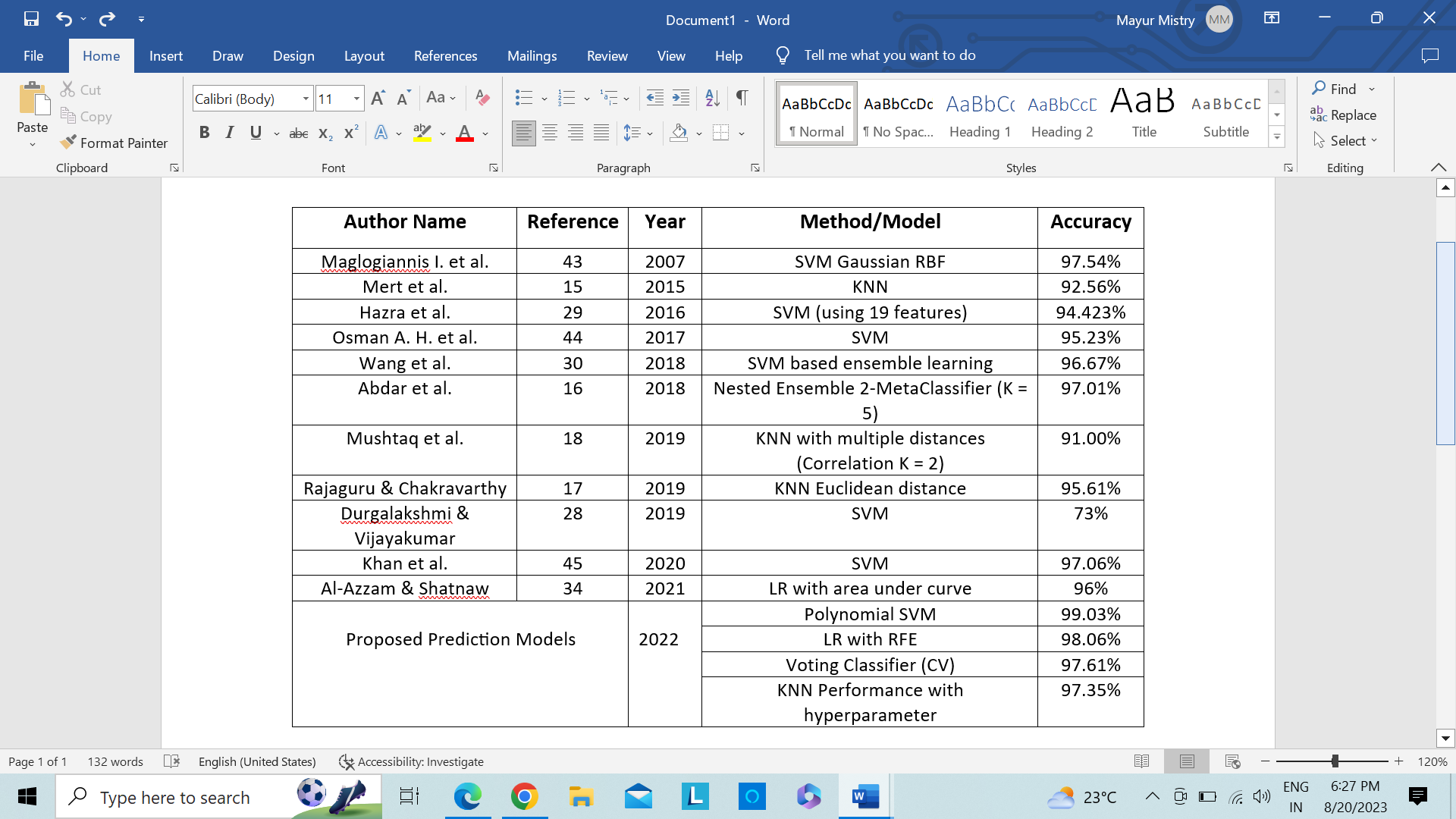


Table 7: Comparison of our suggested BC prediction models' accuracy with that of earlier research that made use of the same WDBC dataset.

# 6.Project Plan and Management

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Total – 6 months | March-23 | April-23 | May-23 | June-23 | July-23 | Aug-23 |
| Literature Review | ✔️ |  |  |  |  |  |
| Mathematical Modeling | ✔️ | ✔️ |  |  |  |  |
| Proposed Framework |  |  | ✔️ |  |  |  |
| Experimentation |  |  |  | ✔️ |  |  |
| Comparison With Existing Approaches |  |  |  |  | ✔️ |  |
| Thesis Writing |  |  |  |  |  | ✔️ |

Table 8: Project Plan and Management

# 7. Discussion

The F1 score was primarily taken into account in our outcomes evaluations while analysing our data. Largely unbalanced class distributions occurred in datasets, much as they do in classification tasks in the real world. The feature selection distribution outcomes reveal some observations that significantly differ between the classes. For instance, the delivery of benign as well as malignant categories differed significantly in the Supplementary Note 01 concavity mean. To balance these traits, the resampling techniques of Cross-validation, undersampling as well as sampling were used. The minority classes are duplicated by the oversampling strategy, however it causes a problem with overfitting for ML algorithms. The undersampling strategy, in contrast, eliminates the popular categoeries that exclude potential data. These drawbacks might reduce ML accuracy for specific issues like fraud detection, facial recognition, disease identification, etc. Due to the issue with cancer detection, we do not use the oversampling or undersampling strategies in this work. The cross-validation technique, however, was recommended by the author (Kuhn, M.; Johnson, K., 2013) as a dominant technique to deal with the unbalanced class distribution. In cross-validation, a model is trained and tested using multiple subsets of the data. In order to balance the positive and negative characteristics in training and testing datasets, this study used the cross-validation technique using the GridSearchCV and the k-fold. Due to the effective handling of real and predicted classes by TP, TN, FP, and FN, recall, precision as well as F1 score are included in the cross-validation matrices, were compared. In Section 3.2, these measures' correct meanings are provided.

Our suggested prediction model successfully recognised the tumour and classified the cancer characteristics as malignant thanks to a 99.3% F1 score in the polynomial Support Vector Machine implementation. Therefore, a greater F1 score correlates with a better ability to diagnose tumours. Table 7 compares the F1 score as well as accuracy of this investigation with earlier studies that used the WDBC dataset. The data analyst would use these predictive algorithms through DM approaches to help find the malignant mass by examining the cancerous data. Similar to Figure 7, Figure 8 shows how models and methodologies' performance can be compared using cross-validation techniques. Because the ML models must also deal with the time complexity, Table 6 gives the execution times of each model with the least amount of precision possible. . Therefore, based on the analysis above, our contribution of the prediction models and approaches we've suggested can effectively aid the cancer domain in obtaining very satisfactory outcomes for the detection of BC.

In this research, the goal of using ML models to detect BC with the best accuracy was achieved. The results may not be comparable to as well as effective through the data from Asian patients because these datasets are from American patients. In relation to the study's weaknesses, which might be overcome in the future by using a new dataset and applying neural networks.

# 8. Conclusion

To ensure good care in the healthcare sector, an accurate and early identification of many diseases, such as BC, remains a significant challenge. Due to the abundance of data and the lack of DM approaches with suitable ML classifiers, the exact study of cancer characteristics remains a time-consuming and difficult task. In this study, four distinct ML prediction models—SVM, LR, KNN, and EC—were proposed with a four-layered essential DET to detect BC tumors and categories into benign as well as malignant types tumors. The application of DET prior to performing ML classifiers on the WDBC datasets was one of the study's main goals. We were able to enhance the prediction model's results with a greatest F1 score as well as greater accuracy score than before thanks to these mining strategies. The key conclusion showed that the initial prediction model had attained the best accuracy (99.3%). DET may successfully detect better accuracy since Recursive features in LR removal also achieved 98.06% accuracy. Our results show how well our prediction models perform in terms of diagnosing BC and deliver appropriate results after only a brief period of model training. The doctor and data analyst might use these advanced models, methodologies, and findings to use a more capable classifier to identify BC characteristics.

We will employ deep learning models to identify breast cancer using unique data augmentation procedures as well as DET to manage the diversity and lack of data as soon as image data related to BC becomes accessible. Future work, we would do tests on datasets from various nations to determine whether the performance of the model is impacted by the patient data from diverse geographic areas.

# 9. References

Abdar, M.; Yen, N.Y.; Hung, J.C.S. (2018). Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J. Med. Biol. Eng., 38, 953–965.

Abdar, M.; Zomorodi-Moghadam, M.; Zhou, X.; Gururajan, R.; Tao, X.; Barua, P.D.; Gururajan, R. (2020). A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit. Lett., 132, 123–131.

[Abunasser, Basem S](https://www.proquest.com/openview/496560430a0455d0e25a4a0bb9a89f02/1?pq-origsite=gscholar&cbl=5444811); [AL-Hiealy, Mohammed Rasheed J](https://www.proquest.com/openview/496560430a0455d0e25a4a0bb9a89f02/1?pq-origsite=gscholar&cbl=5444811); [Zaqout, Ihab S](https://www.proquest.com/openview/496560430a0455d0e25a4a0bb9a89f02/1?pq-origsite=gscholar&cbl=5444811); [Abu-Naser, Samy S](https://www.proquest.com/openview/496560430a0455d0e25a4a0bb9a89f02/1?pq-origsite=gscholar&cbl=5444811). (2022).

[International Journal of Advanced Computer Science and Applications](https://www.proquest.com/openview/496560430a0455d0e25a4a0bb9a89f02/1?pq-origsite=gscholar&cbl=5444811)**;** West Yorkshire[Vol. 13, Iss. 7.](file:///C:\Users\Mayur\Desktop\Jaydeep\ Vol. 13, Iss. 7.%20 ) <https://doi.org/10.14569/IJACSA.2022.0130729>

Akay, M.F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Appl. 2009, 36, 3240–3247.

Al-Azzam, N.; Shatnawi, I. (2021). Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. Ann. Med. Surg., 62, 53–64.

Bicchierai, G.; Di Naro, F.; De Benedetto, D.; Cozzi, D.; Pradella, S.; Miele, V.; Nori, J. (2021). A Review of Breast Imaging for Timely Diagnosis of Disease. Int. J. Environ. Res. Public Health, 18, 5509.

Brause, R.W. (2001). Medical analysis and diagnosis by neural networks. In Proceedings of the International Symposium on Medical Data Analysis, Madrid, Spain, 8–9 October 2001; Springer: Berlin/Heidelberg, Germany, pp. 1–13.

Bunterngchit, C.; Leepaitoon, S. (2019). Simulation-Based Approach for Reducing Goods Loading Time. In Proceedings of the 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain.

Durgalakshmi, B.; Vijayakumar, V. (2020). Feature selection and classification using support vector machine and decision tree. Comput. Intell., 36, 1480–1492.

Durga Prasad Jasti, Abu Sarwar Zamani, K. Arumugam, Mohd Naved, Harikumar Pallathadka, F. Sammy, Abhishek Raghuvanshi, Karthikeyan Kaliyaperumal (2022). "Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis", *Security and Communication Networks*, vol., Article ID 1918379, 7 pages, 2022. <https://doi.org/10.1155/2022/1918379>

Ebrahim, M.; Sedky, A.A.H.; Mesbah, S. (2023). Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer. Data, 8, 35. <https://doi.org/10.3390/data8020035>

Epimack Michael, He Ma, Hong Li, Shouliang Qi, (2022). "An Optimized Framework for Breast Cancer Classification Using Machine Learning", *BioMed Research International*, vol. 2022, Article ID 8482022, 18 pages. <https://doi.org/10.1155/2022/8482022>

Fu, Y.; Jung, A.W.; Torne, R.V.; Gonzalez, S.; Vöhringer, H.; Shmatko, A.; Gerstung, M. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer, 1, 800–810.

Furey, T.S.; Cristianini, N.; Duffy, N.; Bednarski, D.W.; Schummer, M.; Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16, 906–914.

Goudarzi S., J. Whyte, M. Boily, A. Towers, R. D. Kilgour and H. Rivaz, (2023). "Segmentation of Arm Ultrasound Images in Breast Cancer-Related Lymphedema: A Database and Deep Learning Algorithm," in IEEE Transactions on Biomedical Engineering, <https://doi.org/10.1109/TBME.2023.3253646>.

Gupta S., A. Panwar, R. Yadav, M. Aeri and M. Manwal, (2022). "Employing Deep Learning Feature Extraction Models with Learning Classifiers to Diagnose Breast Cancer in Medical Images," IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-6, <https://doi.org/10.1109/DELCON54057.2022.9752856>.

Hager Saleh, Sara F. Abd-el ghany, Hashem Alyami, Wael Alosaimi, (2022). "Predicting Breast Cancer Based on Optimized Deep Learning Approach", *Computational Intelligence and Neuroscience*, vol., Article ID 1820777, 11 pages, 2022. <https://doi.org/10.1155/2022/1820777>

Hanan Aljuaid, Nazik Alturki, Najah Alsubaie, Lucia Cavallaro, Antonio Liotta, (2022). Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning, Computer Methods and Programs in Biomedicine, Volume 223, 106951, ISSN 0169-2607, <https://doi.org/10.1016/j.cmpb.2022.106951>.

Hasan, A.S.M.T.; Sabah, S.; Haque, R.U.; Daria, A.; Rasool, A.; Jiang, Q. (2022). Towards Convergence of IoT and Blockchain for Secure Supply Chain Transaction. Symmetry 14, 64.

Hazra, A.; Mandal, S.K.; Gupta, (2016). A. Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms. Int. J. Comput. Appl., 145, 39–45.

Huang, C.L.; Liao, H.C.; Chen, M.C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Syst. Appl., 34, 578–587.

Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genom. Proteom. 15, 41–51.

Haq A. U. et al., (2022). "DEBCM: Deep Learning-Based Enhanced Breast Invasive Ductal Carcinoma Classification Model in IoMT Healthcare Systems," in IEEE Journal of Biomedical and Health Informatics, <https://doi.org/10.1109/JBHI.2022.3228577>.

Jafarzadeh, H.; Mahdianpari, M.; Gill, E.; Mohammadimanesh, F.; Homayouni, S. (2021). Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation. Remote Sens.

Jabeen, K.; Khan, M.A.; Alhaisoni, M.; Tariq, U.; Zhang, Y.-D.; Hamza, A.; Mickus, A.; Damaševičius, R. (2022). Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion. Sensors, 22, 807. <https://doi.org/10.3390/s22030807>

Jiang, X.; Xu, C. (2022). Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastasis Using Clinical Data. J. Clin. Med., 11, 5772. <https://doi.org/10.3390/jcm11195772>

Kamyab, M.; Tao, R.; Mohammadi, M.H. (2018). Sentiment Analysis on Twitter. In Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality—AIVR 2018, Taichung, Taiwan, 10–12 December.

Khan, F.; Khan, M.A.; Abbas, S.; Athar, A.; Siddiqui, S.Y.; Khan, A.H.; Hussain, M. (2020). Cloud-based breast cancer prediction empowered with soft computing approaches. J. Healthc. Eng., 8017496.

Khandezamin, Z.; Naderan, M.; Rashti, M.J. (2020). Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. J. Biomed. Inform. 111, 103591.

Kirola M, Memoria M, Dumka A, Joshi K. (2022). A Comprehensive Review Study on: Optimized Data Mining, Machine Learning and Deep Learning Techniques for Breast Cancer Prediction in Big Data Context. Biomed Pharmacol J;15(1)

Kuhn, M.; Johnson, K. (2013). Applied Predictive Modeling; Springer: New York, NY, USA.

Leão, D.C.M.R.; Pereira, E.R.; Pérez-Marfil, M.N.; Silva, R.M.C.R.A.; Mendonça, A.B.; Rocha, R.C.N.P.; García-Caro, M.P. (2021). The Importance of Spirituality for Women Facing Breast Cancer Diagnosis: A Qualitative Study. Int. J. Environ. Res. Public Health, 18, 6415.

Lim, J.; Sohn, J.; Sohn, J.; Lim, D. (2013). Breast cancer classification using optimal support vector machine. J. Korea Soc. Health Inform. Stat., 38, 108–121.

Liu L. et al., (2023). "Collaborative Transfer Network for Multi-Classification of Breast Cancer Histopathological Images," in IEEE Journal of Biomedical and Health Informatics, doi: <https://doi.org/10.1109/JBHI.2023.3283042>.

Liu M. et al., (2022). "A Deep Learning Method for Breast Cancer Classification in the Pathology Images," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 10, pp. 5025-5032, Oct., <https://doi.org/10.1109/JBHI.2022.3187765>.

Lopez-Pérez M., P. Morales-Álvarez, L. A. D. Cooper, R. Molina and A. K. Katsaggelos, (2023). "Deep Gaussian Processes for Classification with Multiple Noisy Annotators. Application to Breast Cancer Tissue Classification," in IEEE Access, vol. 11, pp. 6922-6934, doi: <https://doi.org/10.1109/ACCESS.2023.3237990>.

Maglogiannis, I.; Zafiropoulos, E.; Anagnostopoulos, I. (2009). An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Appl. Intell., 30, 24–36.

McWilliam, A.; Faivre-Finn, C.; Kennedy, J.; Kershaw, L.; Van Herk, M.B. (2016). Data mining identifies the base of the heart as a dose-sensitive region affecting survival in lung cancer patients. Int. J. Radiat. Oncol. Biol. Phys., 96, S48–S49.

Mejdoub, M.; Amar, C.B. (2013). Classification improvement of local feature vectors over the KNN algorithm. Multimed. Tools Appl., 64, 197–218.

Mert, A.; Kılıç, N; Bilgili, E.; Akan, A. (2015). Breast cancer detection with reduced feature set. Comput. Math. Methods Med., 265138.

Mondéjar-Guerra, V.; Novo, J.; Rouco, J.; Penedo, M.G.; Ortega, M. (2019). Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers. Biomed. Signal Process. Control, 47, 41–48.

Mushtaq, Z.; Yaqub, A.; Sani, S.; Khalid, A. (2020). Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. J. Chin. Inst. Eng., 43, 80–92.

Muzammal, M.; Qu, Q.; Nasrulin, B. (2019). Renovating blockchain with distributed databases: An open-source system. Future Gener. Comput. Syst., 90, 105–117.

Natalie Baughan, BS and others, (2022). Past, Present, and Future of Machine Learning and Artificial Intelligence for Breast Cancer Screening, Journal of Breast Imaging, Volume 4, Issue 5, September/October, Pages 451–459, <https://doi.org/10.1093/jbi/wbac052>

Nusrat Mohamud din, Rayees Ahmad Dar, Muzafar Rasool, Assif Assad, (2022). Breast cancer detection using deep learning: Datasets, methods, and challenges ahead, Computers in Biology and Medicine, Volume 149, 106073, ISSN 0010-4825, <https://doi.org/10.1016/j.compbiomed.2022.106073>.

Ortega-Martorell, S., Riley, P., Olier, I. *et al.* (2022). Breast cancer patient characterisation and visualisation using deep learning and fisher information networks. *Sci Rep* **12**, 14004. <https://doi.org/10.1038/s41598-022-17894-6>

Osman, A.H. (2017). An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. Int. J. Adv. Comput. Sci. Appl., 8, 158–165.

Park, E.Y.; Yi, M.; Kim, H.S.; Kim, H. (2021). A Decision Tree Model for Breast Reconstruction of Women with Breast Cancer: A Mixed Method Approach. Int. J. Environ. Res. Public Health, 18, 3579.

Park, K.H.; Batbaatar, E.; Piao, Y.; Theera-Umpon, N.; Ryu, K.H. (2021). Deep Learning Feature Extraction Approach for Hematopoietic Cancer Subtype Classification. Int. J. Environ. Res. Public Health, 18, 2197.

Pławiak, P. (2018). Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals. Swarm Evol. Comput., 39, 192–208.

Polat, K.; Güne¸s, S. (2007). Breast cancer diagnosis using least square support vector machine. Digit. Signal Process. 2007, 17, 694–701.

Prasad, Y.; Biswas, K.K.; Jain, C.K. (2010). SVM classifier-based feature selection using GA, ACO and PSO for siRNA design. In Proceedings of the International Conference in Swarm Intelligence, Beijing, China, 12–15 June 2010; Springer: Berlin/Heidelberg, Germany; pp. 307–314.

Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. (2022). Prediction of Breast Cancer using Machine Learning Approaches. J Biomed Phys Eng. Jun 1;12(3):297-308. doi: 10.31661/jbpe.v0i0.2109-1403.

Rajaguru, H. (2019). Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. Asian Pac. J. Cancer Prev. APJCP, 20, 3777.

Rajasekhar D., M. Rafi D, S. Chandre, V. Kate, J. Prasad and A. Gopatoti, (2023). "An Improved Machine Learning and Deep Learning based Breast Cancer Detection using Thermographic Images," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, pp. 1152-1157, <https://doi.org/10.1109/ICEARS56392.2023.10085612>.

Ramesh, S., Sasikala, S., Gomathi, S. *et al.* (2022). Segmentation and classification of breast cancer using novel deep learning architecture. *Neural Comput & Applic* **34**, 16533–16545. <https://doi.org/10.1007/s00521-022-07230-4>

Rasool, A.; Jiang, Q.; Qu, Q.; Kamyab, M.; Huang, M. (2022). HSMC: Hybrid Sentiment Method for Correlation to Analyze COVID19 Tweets. In Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery; Springer International Publishing: Berlin/Heidelberg, Germany; pp. 991–999.

Rasool, A.; Tao, R.; Kashif, K.; Khan, W.; Agbedanu, P.; Choudhry, N. (2020). Statistic Solution for Machine Learning to Analyze Heart Disease Data. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing, Shenzhen, China, 15–17 February; pp. 134–139.

Rahman A., Y. Zhang and J. G. Park, (2023). "A Breast Cancer Recurrence Risk Predictor Using a Deep Learning Multi-Omics Data Integration Framework," 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, pp. 921-922, <https://doi.org/10.1109/CBMS58004.2023.00343>.

Sajiv G., Ramkumar G., (2022). "Multiple Class Breast Cancer Detection Method Based on Deep Learning and MIRRCNN Model," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, pp. 981-987, <https://doi.org/10.1109/ICICT54344.2022.9850707>.

Sammut, SJ., Crispin-Ortuzar, M., Chin, SF. *et al.* (2022). multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629. <https://doi.org/10.1038/s41586-021-04278-5>

Sekar R.R., M. R. Kumar, M. V. Vardhan, K. V. Gopal and C. Jayakkam Allan Tilak, (2023). "Approach to Predict Early Stage of Breast Cancer using Machine Learning," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 122-125, doi: <https://doi.org/10.1109/ICICCS56967.2023.10142851>.

Seddik, A.F.; Shawky, D.M. (2015). Logistic regression model for breast cancer automatic diagnosis. In Proceedings of the 2015 SAI Intelligent Systems Conference (IntelliSys), London, UK, 10–11 November; pp. 150–154.

Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer statistics, (2022). CA Cancer J. Clin. 2022, 72, 7–33.

Soham Chattopadhyay, Arijit Dey, Pawan Kumar Singh, Diego Oliva, Erik Cuevas, Ram Sarkar, (2022). MTRRE-Net: A deep learning model for detection of breast cancer from histopathological images, Computers in Biology and Medicine, Volume 150, 106155, ISSN 0010-4825, <https://doi.org/10.1016/j.compbiomed.2022.106155>.

Subashini, T.S.; Ramalingam, V.; Palanivel, S. (2009). Breast mass classification based on cytological patterns using RBFNN and SVM. Expert Syst. Appl., 36, 5284–5290.

Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin., 71, 209–249.

Sushovan Chaudhury, Alla Naveen Krishna, Suneet Gupta, K. Sakthidasan Sankaran, Samiullah Khan, Kartik Sau, Abhishek Raghuvanshi, F. Sammy, (2022). "Effective Image Processing and Segmentation-Based Machine Learning Techniques for Diagnosis of Breast Cancer", *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 6841334, 6 pages. <https://doi.org/10.1155/2022/6841334>

Tolles, J.; Meurer, W.J. (2016). Logistic regression: Relating patient characteristics to outcomes. JAMA, 316, 533–534.

Tiesto D. et al., (2023). "Multi-Input Deep Learning Approach for Breast Cancer Screening Using Thermal Infrared Imaging and Clinical Data," in IEEE Access, vol. 11, pp. 52101-52116, doi: <https://doi.org/10.1109/ICEARS56392.2023.10085612>

Wang, H.; Zheng, B.; Yoon, S.W.; Ko, H.S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur. J. Oper. Res., 267, 687–699.

Wang, X.; Ahmad, I.; Javeed, D.; Zaidi, S.A.; Alotaibi, F.M.; Ghoneim, M.E.; Daradkeh, Y.I.; Asghar, J.; Eldin, E.T. (2022). Intelligent Hybrid Deep Learning Model for Breast Cancer Detection. Electronics, 11, 2767. <https://doi.org/10.3390/electronics11172767>

Xueheng Liang, Xingyan Yu, Tianhu Gao (2022). Machine learning with magnetic resonance imaging for prediction of response to neoadjuvant chemotherapy in breast cancer: A systematic review and meta-analysis, European Journal of Radiology, Volume 150, 110247, ISSN 0720-048X, <https://doi.org/10.1016/j.ejrad.2022.110247>.

Yang, L.; Xu, Z. (2019). Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning. Int. J. Mach. Learn. Cybern, 10, 591–601.

Yash Amethiya, Prince Pipariya, Shlok Patel, Manan Shah (2022). Comparative analysis of breast cancer detection using machine learning and biosensors, Intelligent Medicine, Volume 2, Issue 2, Pages 69-81, ISSN 2667-1026, <https://doi.org/10.1016/j.imed.2021.08.004>.

Yu, Z.; Chen, H.; Liu, J.; You, J.; Leung, H.; Han, G. (2015). Hybrid k-nearest neighbor classifier. IEEE Trans. Cybern., 46, 1263–1275.

Zheng, B.; Yoon, S.W.; Lam, S.S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. 41, 1476–1482.

Zourhri M., S. Hamida, N. Akouz, B. Cherradi, H. Nhaila and M. E. Khaili, (2023). "Deep Learning Technique for Classification of Breast Cancer using Ultrasound Images," 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Mohammedia, Morocco, pp. 1-8, doi: <https://doi.org/10.1109/IRASET57153.2023.10153069>.