Investigating The Roles of microRNAs / lncRNAs in Characterizing Breast Cancer Subtypes and Prognosis

Reyhan Zeynep PEK[1] Muhammed Talha ZAVALSIZ[1] Melis SERDAR[1] Lama Alhajj[2] Sleiman Alhajj[2]
Kashfia Sailunaz[3] Tansel Özyer[4] Jon Rokne[3] Reda Alhajj [1,3,5]
[1]Department of Computer Engineering, Istanbul Medipol University, Istanbul, Turkey
[2]International School of Medicine, Istanbul Medipol University, Istanbul, Turkey
[3]Department of Computer Science, University of Calgary, Alberta, Canada
[4]Department of Computer Engineering, Ankara Medipol University, Ankara, Turkey
[5]Department of Health Informatics, University of Southern Denmark, Odense, Denmark

Abstract

Molecular subtyping is a method of separating tumor clusters in a cancer type with common features according to molecular data and classification models. Genome datasets are taken from many different people and some genetic material, more precisely genetic markers, are obtained to predict the presence of a disease. In addition, breast cancer occurs due to mutation or modification observed in cells. miRNAs and lncRNAs take participation in cell cycle, regulation, and even chromatic inhibition of cell. For example, miRNAs function in cell cycle regulation as the degradation of mRNAs. Therefore, the aim of this work is to investigate the roles of miRNAs and lncRNAs in prognosis and characterizing the subtypes of Breast Cancer. ***Index Terms*—lncRNA, miRNA, Breast Cancer, Subtyping, Python, Machine Learning Models, SVM, DT, NB, Adaboost, RF, MLP, LR, SGD, KNN**

I. INTRODUCTION

Cancer refers to a disease that is caused by uncontrollable growth and spread of cells while making other healthy cells unable to function. And cancer subtyping is an important term to distinguish the subtypes since the effectiveness of treatments depends on more personalised treatments. And for breast cancer subtypes which demonstrates different behaviours and features according to their genetic background, general treatment methods are not efficient for each subtype. Even the subtyping is defined with behaviours and features and the places that cancer is observed, their genomic background similarities and differences leads them to be defined as the subtypes. Which means a modification or mutation in genome can be the reason of breast cancer and their subtypes.

As the genome, RNA specifies an important role in cell life such as in cell growth, cell cycle, synthesis of proteins and control of regulations. micro RNAs (miRNAs) and long noncoding RNAs (lncRNAs) are part of RNAs, and they take place in those important roles of RNA. According to the studies it is predicted that microRNAs and long noncoding RNAs shows similarities for the subtypes of breast cancer. However, in vitro and in vivo micro RNAs and long noncoding RNAs are highly incompatible for the examinations. Therefore, with the databases of micro RNAs and long noncoding RNAs the relation between breast cancer subtypes and those RNA subtypes can be investigated computationally which is the purpose of the work described in this paper.

Firstly, we need to find a dataset that includes data from breast cancer patients and where we can analyze microRNA/lncRNAs. We may obtain this dataset from references to relevant articles or from other sites that publish datasets. Among the features that the data set should include, breast cancer patients' microRNA expressions and lncRNA expressions should be included. We can analyze the roles of microRNA/lncRNA expressions for breast cancer subtyping and prognosing using the methods determined after the data set is obtained. After the computational methods to be used are determined, the results of these methods can be compared, and it can be observed which method gives better results. Various metrics are planned to be used to evaluate these methods. Progress of the plan:

1) Find Dataset including breast cancer patients with microRNA/lncRNA data expressions.
2) Determine the exact methods to be used in this project.
3) Analyze the dataset with specified methods.
4) Pre-process the Dataset if it is needed.
5) Measure the performance of methods with various metrics.
6) Compare the results and understand the role of microRNA and lncRNAs in characterizing breast cancer subtypes and prognosis.

II. BACKGROUND

A. RNA

RNA or ribonucleic acid refers to a molecule analogue to DNA (deoxyribonucleic acid), yet with a single strand. RNA single strand transcription is observed from DNA target region (gene) through 3'-5' direction, which is specified according to the synthesized protein and observed as 5' to 3' direction. RNA strand consists of four different bases nucleotides, bases are adenine (A), uracil (U), cytosine (C), and thymine (T). Mainly, to produce protein by strands three types of RNA exists as

messenger RNA shortly mRNA, ribosomal RNA shortly rRNA, and transfer RNA shortly tRNA. However, in recent days, petty RNAs have been discovered to be included in the gene expression regulation. These petty endogenous RNA classes include small transfer RNA, small nuclear RNA, small interfering RNA, microRNA, and lncRNA [2][3].

B. MicroRNA

MicroRNAs which are expressed also as miRNA, refer to petty, quite conversed, single – stranded, non – coding RNA particles included in the gene expression regulation, by binding target mRNA to restrain protein production. In general, miRNAs are consisted of 18 – 24 nucleotides long. miRNA is proceeded from double – stranded 60 to 70 nucleotide RNA region. MiRNAs are usually found in junk DNA which refers to historically intergenic regions and introns of protein coding genes inside genome, where their function is unknown. [2][4][5]

MicroRNA synthesis starts with its gene, after coding from microRNA gene primary microRNA shortly pri– miRNA is observed by transcription. Cleavage of primary microRNA designates precursor microRNA (pre - miRNA), and cleavage of pre – miRNA, microRNA duplex is obtained and finally unwinding RISC assembly mature microRNA is procured, accordingly to their target recognition, they achieve their objectives as gene silencing. The function of microRNA is studied and observed, yet biogenesis and gene silencing of microRNAs are ambiguous [2].

Even though it is uncertain, miRNA mediated silencing of gene is observed by miRNA induced silencing complex or shortly miRISC. MiRNA response elements (MREs) refer to the target mRNA complementary sequences interaction with miRISC where this and miRNA complementary interaction degree penetrates AGO2 endonuclease activity, which is RNA silencing endonuclease, and specifies mRNA cleavage [5]. If the base pairing (interaction) between miRNA and mRNA AGO2 is inhibited and slicer – independent gene silencing mechanism takes place. Furthermore, the major data demonstrates miRNA mediated gene silencing is observed to be in cytoplasm and P – bodies. Also, P – bodies which refer to dynamic small cytoplasmic protein spheroid fields discovered in cells, include mRNA, kinds of enzymes, factors that are essential for the process like mRNA decapping, deadenylating, RNA degradation and repression of translation [2].

Most encountered cases show that, $3'$ untranslated region of messenger RNA target region to conduce to the degradation of mRNA and repression of translation. Messenger RNA is transcribed in nucleus from DNA and moved into cytoplasm where mRNA interacts with ribosome to synthesize proteins [6]. MiRNAs possess the activation of translation or transcription maintenance under specific conditions. Interaction of miRNAs with the objected region can be affected from many considerations such that subcellular location, or abundancy of miRNAs and target messenger RNAs, also the affinity of miRNA and messenger RNA coaction. miRNA can be secreted into extracellular fluids transferred into destination cells by vesicles which can be exosomes, proteins (with binding to them). MiRNA is functioned in extracellular as chemical messengers for the mediated transmission of cell to cell. On the other hand, extracellular miRNA is considerably denounced as potential biomarkers for a range of diseases since they function as cell to cell mediated transmission. Extracellular miRNA which are highly stable, and resist to degradation, uptake pathway is not understood clearly where the pathway is offered to be vesicle associated entering via phagocytosis, endocytosis, or plasma membrane's direct fusion, otherwise vesicle – free secretion of miRNAs can be completed via significant cell surface receptors [5].

C. Long non-coding RNA

RNA transcription is completed by DNA which has more than three billion base pair, and 70% of the genome can be transcribable to RNA where only 2% of the RNA can be utilized as a protein transcription template. Those possible coding templates are denominated as coding RNA, while other non – available to code RNA referring to non – coding RNA. And, long non – coding RNA shortly known as lncRNA which contains more than 200 nucleotides, refers to a large and various class of RNA molecules. Studies discovered that lncRNA functions in various biological operations such that to combine protein complex, to maintain coding RNA transformation, genomic imprinting, dosage compensation. Division into two major classifications of long non – coding RNA functions as assemble of proteins and competition to bind with non – coding RNAs [7][11].

Long non – coding RNAs can be occurred from entrant particular initiation regions and can endure multiple post – transcriptional alterations. Inside the nucleus, LncRNA is synthesized generally by polymerase II, and deux of them by polymerase III. Posttranscriptional operations can be completed on lncRNAs such like RNA splicing, polyadenylation, and capping which may bring isoforms of lncRNAs from identical progenitor, which rises long non – coding RNAs diversity, and versatility of operation, also processes on lncRNAs can increase stability. Moreover, a significant function that specifies lncRNAs is the requirement of assumption a certain secondary and tertiary structure to bring their function which identifies them from mRNA and miRNA since their function is dependent to on primary sequence, while for lncRNA the interaction depends on tertiary configuration which refers to loss of tertiary structure, causes the loss of function [8][11].

In general, four types of regulatory mechanisms of lncRNAs as epigenetic and transcriptional have been found. The types include lncRNAs interaction firsthand with transcriptional considerations and incline allosteric alteration for the activation with transcriptional elements, another type is when lncRNA behaves like bait via titration of transcription agents detract from chromatin. LncRNA can behave like guide via recruiting modification of chromatin enzymes to specify genes, also, lncRNA can combine multiple proteins to observe ribonucleoprotein (RNP) blocks while behaving like a scaffold. On the other hand, eight types as post – transcriptional regulatory mechanisms have been discovered. The types include, antisense lncRNAs can constitute duplex which can behave as the sense of pre – mRNA, another regulatory mechanism, binding to intron and exon boundary interspace sites of pre – mRNA, and restrain alternating splicing which refers the function that lncRNAs can be correlated to influence splicing and factors of splicing.

Furthermore, lncRNA can comprise hairpin structure that can increase pre – miRNA, and it can comprise the acquaintance site for operational miRNAs, which leads them to be the target of miRNAs. Also, as mentioned before, lncRNA may compete with miRNA for the binding site, if the block of miRNA is successive, mRNA translation will rise. And lastly, interaction between mRNA and lncRNA can form a double – stranded structure which may steer exosome mediated degradation of RNA [8].

D. Breast Cancer

Cancer cells demonstrate similar organisms to healthy cells with DNA and RNA, which is one of the reasons they cannot be detected by the immune system. According to a modification or mutation in DNA and / or RNA which can be observed due to increase of entropy, nuclear radiation, electromagnetic radiation etc., cancer cells are comprised. And when the cancer occurs at breast, it is denominated as breast cancer which is the most common cancer type encountered in women. On the other hand, breast cancer contains four stages, starting from 0 to IV [12].

Breast is comprised of two types of tissues as glandular and stromal or known as supporting tissues. Glandular tissue contains milk generating glands known as lobules and milk passages known as ducts, thus stromal tissue containing fatty and fibrous connective tissues. Moreover, lymphatic tissue refers to breast immune system tissue which cleans cellular waste and fluids. Even there are two tissue types, several types of breast cancer can be developed according to their areas [12].

The types of breast cancer:

- According to site of the cancer -
  - Non – invasive breast cancer: When the cancer cells are comprised around the ducts and not have any interaction with surrounding fatty and uniting tissue of breast [12].
  - Invasive breast cancer: Type of cancer that cancer cells tear the duct and lobular wall and interact with its surrounding fatty and uniting tissue of breast [12].
- Frequently observed breast cancer -
  - Lobular carcinoma in situ (LCIS): Not spread where it is initially composed, which is milk glands of breast [12].
  - Ductal carcinoma in situ: Not spread where it is initially composed, which is ducts of breast [12].
- Infiltrating lobular carcinoma (ILC): Developing in lobules and often spreading over the body which referring to metastasizes [12].
- Infiltrating ductal carcinoma (IDC): Developing in ducts and often metastasizes [12].
- Less commonly observed breast cancer
  - Medullary carcinoma: Invasive cancer type which comprises of a verge between normal and tumor tissue [12].
  - Mutinous carcinoma Also, known as colloid carcinoma, mutinous carcinoma that is composed of the mucus generating cancer cells [12].
  - Tubular carcinoma A specific type of invasive breast carcinoma [12].
- Inflammatory breast cancer: The cancer type causes an appearance change in inflamed breasts as red and warm with dimples, also can cause thick ridges are formed by cancer cells preventing lymph channels or vessels on the skin of breast which is an excessively rapid growing cancer type [12].
- Paget's disease of the nipple: Developing from milk ducts and metastasizes onto skin of nipple and areola [12].
- Phyllodes tumor: This cancer type can be expressed as non – cancerous or cancerous. In the type tumors start to comprise in uniting tissues of breast [12]. Cell cycle regulation contains stages and check points for a healthy cell with growth factors, yet in cancer cells uncontrolled growth and spread (division of cell through uncontrolled cell cycle) is occurred. Any mutation, or modification in cell cycle can cause cancer, where the cycle cannot be stopped or checked due to a protein synthesize or not being able to synthesize the protein and / or provoke specific cell cycle agents.

In recent studies, long non – coding RNA and miRNA have been become an interesting and dilatational focus of cancer genomic research, lncRNAs relation with diseases is investigated and it has been found that certain types of cancers, some specific lncRNA expression rises [7]. On the other hand, miRNAs are investigated for the regulation of cell cycle and improvement which makes them a prior subject for the investigations of cancer diseases [13].

E. Breast Cancer Subtypes

Breast cancer, which is the most common type of cancer and also the most common cause of death in women, arises with the uncontrolled proliferation of cells in the breast tissue. Breast cancer does not refer to a single disease, it has a wide range of pathological and clinical outcomes and processes due to its many different types. Due to the difference in biological properties, it is of great importance to increase therapeutic strategies and to be target-based in the treatment of breast cancer. Subtype classification is important at this point, and the treatment process can be carried out more effectively and accurately with subtype classification. Studies have shown that the molecular size underlying the disease rather than the prognosis is the main factor in the tumor response [21].

Although there are studies that involve difficulties in the prevention and treatment of breast cancer, it is sometimes difficult to apply these preclinical studies to the clinic. This is due to the nature of breast cancer, unlike other cancer types. Breast cancer is a heterogeneous group of diseases that exhibit different phenotypes and molecular appearances. The breast cancer receptor result [estrogen (ER) and progesterone (PR) hormone receptors, HER-2 (human epidermal growth factor receptor 2)] treatment protocol is determined. Luminal A; Patients with positive/exhibiting ER and PR receptors, Luminal B; It includes ER and PR receptors positive/exhibiting and higher histological grade than Luminal A or triple negative (ER, PR and HER-2 negative/non-exhibiting) group. These subgroups allow us to predict the clinical behavior of the disease, including life expectancy, mode of metastasis, and response to treatment. Within the framework of these data, the appropriate model should be selected for preclinical study [22].

Computational methods, particularly machine learning, have been applied to cancer detection and diagnosis using miRNAs as biomarkers. As an example, one study used hierarchical clustering in 73 bone marrow samples and determined that miRNA expression distinguishes different subtypes of tumors in acute lymphoblastic leukemia. In another study, a k-NN classifier was constructed using lung cancer samples and healthy mouse lung samples, achieving a classification accuracy of 100%. Also in another study, a miRNA-based tissue classifier was created to determine the source location of metastatic tumors and using k-NN and decision trees to classify tumors according to 22 different tumor classes, they achieved 89% accuracy in the validation set. However, previous studies were limited by the amount of miRNA data currently available, and the number of miRNAs known at the time. The amount of miRNA expression data has increased dramatically with the advent of the Genomic Data Commons (GDC) Data Portal provided by the National Cancer Institute. [23,24].

Researchers have been applying different models on miRNA and lncRNA data for breast cancer analysis recently [28, 29]. Søkilde et. al [25] proposed a breast cancer molecular subtype classification model with miRNA expressions using nearest-centroid model. Their survival analysis showed that high expression of miRNA cluster can represent higher survival rate for Luminal A tumors. Another similar study in [26] showed that miRNA can be useful in early breast cancer detection. A recent survey on machine learning methods used in miRNA-based breast cancer analysis was presented in [31]. They chose 6 out of 36 publications collected from PubMed, IEEE, Google Academico, and ScienceDirect between 2018 and 2022. They discussed the methods used in those research (i.e., SVM, RF, ANN, SMO, ANN, KNN, DT, DISCR, and their variations) with the conclusion that SVM and RF are the 2 most popular machine learning models in this type of research work. They also showed that these models are mostly used for breast cancer subtype classification and/or cancer vs. healthy tissue detection with more than 90% and more than 70% accuracy rate, respectively.

Sarkar et. al [27] on the other hand provided a machine learning based breast cancer subtype classification using miRNA data. They applied SVM, ANN, KNN, DT, RF, NB, DISCR on the dataset and chose RF as it performed best and then used feature selection and ensemble ranking for cancer type classification, survival analysis, network analysis, pathways, etc. Andreini et. al [30] proposed another machine learning based approach where an SVM classifier was used on miRNA data to classify them into tumorous and healthy classes. Then an RF classifier was applied only on the tumorous data to classify them further into Basal-like, HER2-enriched, Luminal A and Luminal B subtype classification. The tumor-healthy classification achieved more than 95% accuracy whereas the subtype classification got 78% accuracy. The existing researches on this topic provided a clear idea regrading the current scopes of improvements in breast cancer and breast cancer subtype classifications with miRNA and lncRNA data.

III. METHODOLOGY

A. Data Collection

One of the important aspects of the project is to obtain the dataset. Therefore, dataset should include breast cancer patients' microRNA expressions and lncRNA expressions. In this project, the dataset containing miRNA expressions was provided by Rincon et al. (2019) was obtained for experiments. In the study of the researchers, they obtained and analyzed datasets containing miRNA molecules for different types of cancer. In addition, they analyzed the breast cancer molecular subtype from the GEO and TCGA datasets using different methods [19]. In this project, the dataset obtained from the TCGA data set shared by the researchers and labeled with the subtypes of 764 samples was used. This dataset contains molecular subtype of Breast Cancer for miRNA molecules. Each subtype in the dataset is labeled as a number. There are a total of 5 subtypes in the Label set. These subtypes are Normal, LumA, LumB, Her2 and Basal.

B. Building Machine Learning Models

The intended scientific contribution to this subject is to make analysis using various machine learning methods, different from the methods specified in the literature. By building machine learning models, the roles of microRNAs and lncRNAs in identifying breast cancer subtypes and prognoses can be analyzed. The Python programming language and the scikit-learn library were used to perform the analysis and obtain the results. In this study, nine different machine learning algorithms were used. These are AdaBoost Classifier, Random Forest, Support Vector Machine, Decision Tree, Naïve Bayes, MLP Classifier, Logistic Regression, Stochastic Gradient Descent Classifier (SGD) and K-Nearest Neighbors Classifier. In addition, stratified k-fold cross validation method was used to separate the dataset into training and test sets. With this method, each different data type can be included in the training and test dataset, and this method can increase the prediction performance of the models. For stratified k-fold cv, the number of folds was set to 10. The performance of the models was measured using various metrics.

First of all, the performances of the created models were measured on the dataset containing miRNA molecules expressions. The aim here is to observe whether the models correctly predict the label set or not. There are varieties of breast cancer subtypes in the label set so that it can be observed whether miRNA is useful in predicting breast cancer subtypes.

C. Hyper-parameter optimization

The first results of the models were obtained using the default parameters. Thus, after parameter adjustments, the effect of hyper-parameter optimization on the performance of the models can be compared. Although the performance of some models was low for the first results, it was expected that the performance of these models would increase with hyper-parameter optimization. Hyper-parameter optimization can be done using different methods. Some of these methods are Grid Search and Randomized

TABLE I. Table 1: Confusion matrix.

Search techniques. These techniques can be used via the scikit-learn library. These parameter optimization techniques test the models with the specified parameters and generate the model parameters that provide the best performance.

Thus, parameter adjustments of the models are made. Grid search technique tests the model using all the specified parameter sets and shows the parameters that provide the best results. Randomized Search technique makes random parameter combinations and tests the model with random combinations. It shows the parameters that achieve the highest performance. The Grid Search technique makes a lot of combinations using all the parameters, but it is computationally expensive because it takes a lot of time. The Randomized Search technique is a less expensive method as it randomly generates combinations. It can produce a result close to the best performance in less time. In this project, these methods from the scikit-learn library will be used for hyperparameter optimization.

D. Evaluation

Various metrics were used to evaluate the performance of the created models. The metrics used to measure the performance of the models are accuracy, precision, recall and f1 score. These metrics were applied to the models created using the data set containing miRNA expressions and breast cancer subtypes, then performance results were obtained. Using many metrics, we can observe how accurate the prediction made by the model is. Accuracy gives the ratio of correctly estimated samples. Precision gives how many positive samples the model estimates as positive are also positive in the ground truth. Recall gives the result of how many of the samples that the model should predict positive, correctly predicted as positive. The F1 score gives the performance value by taking the harmonic mean of the recall and precision values for the measurement. In this way, it allows us to better compare the performance of models with different precision and recall values. The performance of the models was measured and compared with the specified metrics. In addition, the confusion matrix of several models with good performance results was drawn. The confusion matrix is one of the performance metrics for classification problems. With this table, we can visualize how accurately the model estimates. An example image of the confusion matrix appears in Table 1. The formulas of the metrics used are given next.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
$Precision = \frac{T}{TP+FP}$
$Recall = \frac{TP}{TP+FN}$
$F!Score = \frac{2Recall \times Precision}{Recall+Precision}$

IV. RESULTS AND DISCUSSION

A. miRNA expression data

The dataset includes miRNA molecule expressions and breast cancer subtypes. There is a total of 764 labeled data. The subtypes and their numbers in the dataset are listed in Table 2.

TABLE II. Table 2: Molecular subtypes with miRNA expression of breast cancer patients in the dataset.

Subtypes	miRNA
Normal	33 samples
Lum A	399 samples
Lum B	139 samples
Basal	135 samples
Her2	58 samples

TABLE III. Table 3: Performance values measured with different metrics obtained from the models before parameter tuning.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
SVM	73,7%	74,9%	73,7%	69%
NB	43,4%	39,4%	44%	45%
RF	76,4%	67,6%	76,4%	70,2%
AdaBoost	64,4%	60,6%	64,4%	61,7%
Decision Tree	64,7%	65,5%	65,7%	65,1%
MLP	76,3%	76,8%	76,3%	75,4%
LR	71%	76%	71%	70,8%
SGD	80,2%	81,5%	80,2%	79,3%
KNN	56,5%	40,3%	56,5%	43,9%

First, models were created with a dataset containing breast cancer patients' miRNA expressions and subtypes. Machine learning algorithms SVM, Naïve Bayes, Random Forest, AdaBoost, Decision Tree, MLP, Logistic Regression, Stochastic Gradient Descent and KNN algorithms were used. For the first models created, default parameters were used, and the performances of the models were measured with various metrics. The performance results obtained before parameter tuning are given in Table 3. The performances of the models were measured with accuracy, precision, recall and F1 score metrics. The model with the highest performance before parameter optimization is SGD (Stochastic Gradient Descent) and its accuracy is 80.2%. When Precision, Recall and F1 score metrics are also used, the model with the best performance is the SGD model. It performed quite well before the hyperparameter optimization was done.

TABLE IV. Table 4: Performance values measured with different metrics obtained from the models after parameter tuning

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
SVM	79%	78,3%	79%	77,7%
NB	64,5%	50,2%	50,5%	62%
RF	77,6%	68,1%	77,6%	71,7%
AdaBoost	76,32%	72,89%	76,32%	73,2%
Decision Tree	71%	68,6%	71%	66,1%
MLP	82,9%	82,8%	82,9%	82%
LR	81,6%	81,1%	81,6%	80,8%
SGD	86,8%	87,4%	86,8%	85,5%
KNN	60,5%	65,6%	60,5%	50,4%

TABLE V. Table 5: Parameter settings where models perform best after hyperparameter optimization.

Models	Parameters
GaussianNB	var_smoothing= 0.533669923120631
RandomForestClassifier	criterion='entropy', max_depth=7, max_features='auto', n_estimators=25
SVC	kernel= 'linear', gamma= 0.001, degree= 6, decision_function_shape= 'ovo', C= 100
AdaBoostClassifier	n_estimators= 500, learning_rate= 0.1, algorithm= 'SAMME'
DecisionTreeClassifier	splitter= 'best', min_samples_leaf= 10, max_depth= 3, criterion= 'gini'
MLPClassifier	solver= 'adam', learning_rate= 'constant', hidden_layer_sizes= (50,100,50), alpha= 0.0001, activation= 'relu'
LogisticRegression	C= 0.12016539538694185, max_iter= 300, multi_class= 'auto', penalty= 'none', solver= 'saga'
SGDClassifier	penalty= 'l2', max_iter= 150, loss= 'perceptron', alpha= 0.0001
KNN	n_neighbors=10, metric= 'minkowski', leaf_size= 100, algorithm= 'brute', weights='distance'

B. Before parameter optimization

Hyperparameter optimization was done with Randomized Search technique. As a result, the parameters that gave the best performance were used to create the models. The performances of the models obtained after parameter tuning are given in Table 4. Looking at the performance results of the models, it is observed that the SGD (Stochastic Gradient Descent) model is again the model with the highest performance. This model achieved 86.8% accuracy performance value. In addition, looking at the 80.2% accuracy value obtained from the SGD model before parameter tunning, it was observed that the performance of the model increased to 86.8% after parameter tunning. In addition, the MLP and LR models are other top-performing models. Accuracy values were 82.9% and 81.6%, precision values were 82.8% and 81.1%, recall values were 82.9% and 81.6%, and F1 score values were 82% and 80.8%, respectively. If the results of the precision, recall and F1 score metrics of the SGD model are examined, the values of 87.4%, 86.8% and 85.5% were obtained, respectively. Compared to the performances of other models, the model with the highest performance is SGD.

Thus, it can be said that SGD, MLP and LR models can predict breast cancer subtypes with high performance in the dataset containing miRNA expressions. Also, it is concluded that hyperparameter optimization (shown in Table 5) improves the performance of all models. For this reason, making hyperparameter optimization for the created models and using the appropriate parameters for each model is an important factor for the prediction performance of the models.

C. After parameter optimization

Confusion matrices of the four models that provide the best performance are included in Figures 1, 2, 3 and 4. With these confusion matrices, it can be observed how accurately the models

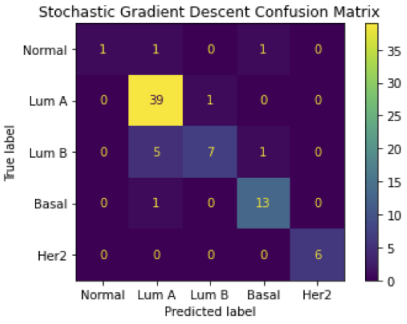

Fig. 1. Figure 4: Stochastic Gradient Descent model confusion matrix.

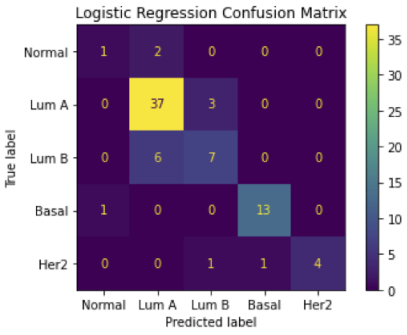

Fig. 2. Figure 5: Logistic Regression model confusion matrix.

predicted each subtype. The four models that provide the best performance are the Stochastic Gradient Descent, MLP Classifier, Logistic Regression, and SVM.

Furthermore, bar plots were drawn based on the performance scores of each model when evaluated with accuracy, precision, recall and f1 score metrics. These bar plotting were done to visualize the comparison of performance results when the models were measured with different metrics. Bar plottings can be seen from figures 5, 6, 7 and 8. It is observed that the model providing the highest performance is SGD.

V. CONCLUSION

Breast cancer is the most common type of cancer in women and has a high fatality rate. According to the latest studies, it has been observed that the treatments applied according to the anatomical markers of this disease create an incorrect process for the patient, and a more detailed investigation of the disease has been carried out. According to the results obtained, Breast Cancer is a type of cancer that should be addressed at the molecular level, and research has focused on this problem. Accordingly, the disease is divided into some groups, and it is aimed to treat the cases in the most accurate way according to their molecular structure and nature. With this project, it was aimed to support the mentioned cases, it was aimed to make a miRNA-based

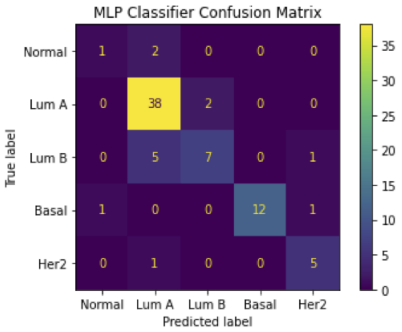

Fig. 3. Figure 6: MLP Classifier model confusion matrix.

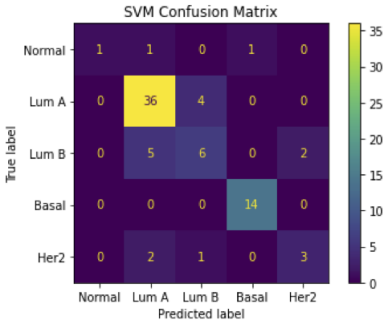

Fig. 4. Figure 7: SVM model confusion matrix.

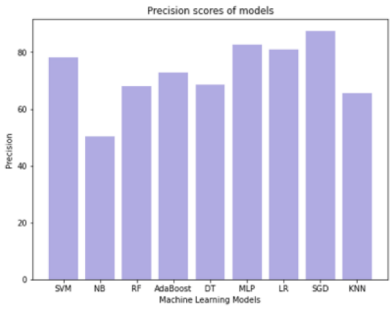

Fig. 6. Figure 9: Comparison of precision scores of all models with bar plotting.

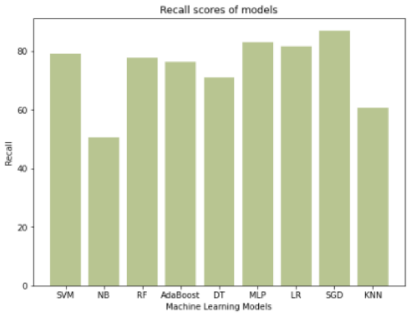

Fig. 7. Figure 10: Comparison of recall scores of all models with bar plotting.

classification using various machine learning techniques, and in this way, to determine the Breast Cancer subtype with a high success rate and to use this information in researches to be made by joining the literature and in breast cancer medical processes is one of the main goals. has been determined. After obtaining the datasets, it was aimed to create targeted algorithms and thus to make Breast Cancer subtype classification.

In this work, 9 machine learning algorithms were used. These algorithms are SVM, Naive Bayes, Random Forest, AdaBoost, Decision Tree, MLP, Logistic Regression, SGD and k-NN. Since the hyperparameter tuning process was not performed in the first models created, the desired results could not be fully obtained when the evaluation metrics were examined. The evaluation metrics used are accuracy, precision, recall and f1 score. The accuracy value alone can sometimes be lacking in evaluating the model, so other evaluation metrics were also applied, and their results were obtained. Hyperparameter tuning is a process that increases the performance of the created models and provides more meaningful results. It is different for each algorithm and

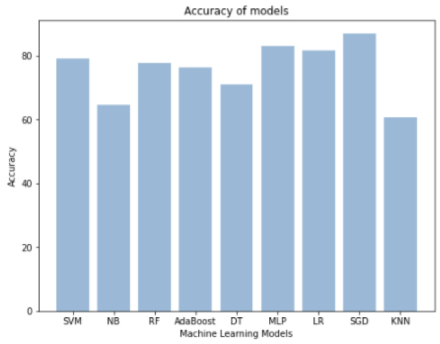

Fig. 5. Figure 8: Comparison of accuracy scores of all models with bar plotting.

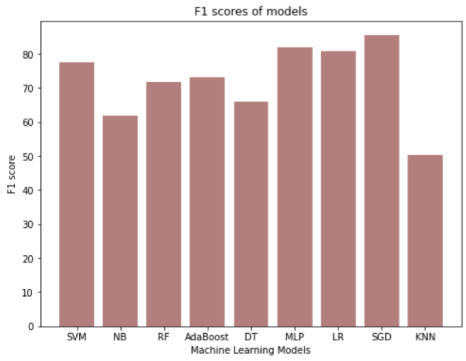

Fig. 8. Figure 11: Comparison of f1 scores of all models with bar plotting.

hyperparameters vary according to the created model. Considering these, hyperparameter tuning processes were carried out. Before this process, the SGD model is the model with the highest performance, while the Naive Bayes model is at the lowest level in terms of success rate compared to other models.

Hyperparameter tuning was performed using the Randomized Search technique, and SGD was the model with the highest success rate. In addition, MLP and LR models were also ranked 2nd and 3rd in terms of success rate. The success rates have increased with the Hyperparameter tuning process and the effect of this phenomenon on Machine learning models has been revealed. The models created have performed the Breast Cancer Subtype Classification process with the success rates in Table 4, and good results have been obtained in general. SGD has been included as the most efficient model in this project and has proven to be a good candidate to be used in this and similar projects.

References

[1] Li X, Truong B, Xu T, Liu L, Li J, Le TD. Uncovering the roles of microRNAs/lncRNAs in characterising breast cancer subtypes and prognosis. BMC Bioinformatics. 2021;22(1):1-22. doi:10.1186/s12859-021-04215-3

[2] MacFarlane L-A, Murphy PR. MicroRNA: Biogenesis, Function and Role in Cancer. Curr Genomics. 2010;11(7):537. doi:10.2174/138920210793175895

[3] Ribonucleic Acid (RNA). https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid. Accessed November 9, 2021.

[4] microRNAs - function & biogenesis - What are miRNAs? https://www.tamirna.com/micrornas-function-biogenesis/. Accessed November 9, 2021.

[5] O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front Endocrinol (Lausanne). 2018;9(AUG):402. doi:10.3389/FENDO.2018.00402/BIBTEX

[6] Messenger RNA (mRNA). https://www.genome.gov/genetics-glossary/messenger-rna. Accessed November 9, 2021.

[7] Morlando M, Ballarino M, Fatica A. Long non-coding RNAs: New players in hematopoiesis and leukemia. Front Med. 2015;2(APR). doi:10.3389/FMED.2015.00023

[8] Yang G, Lu X, Yuan L. LncRNA: A link between RNA and cancer. Biochim Biophys Acta - Gene Regul Mech. 2014;1839(11):1097-1109. doi:10.1016/J.BBAGRM.2014.08.012

[9] Kung JTY, Colognori D, Lee JT. Long Noncoding RNAs: Past, Present, and Future. Genetics. 2013;193(3):651. doi:10.1534/GENETICS.112.146704

[10] Zampetaki A, Albrecht A, Steinhofel K. Long non-coding RNA structure and function: Is there a link? Front Physiol. 2018;9(AUG):1201. doi:10.3389/FPHYS.2018.01201/BIBTEX

[11] Lesizza P, Paldino A, Merlo M, Sinagra G, Giacca M. Noncoding RNAs in Cardiovascular Disease. Nucleic Acid Nanotheranostics Biomed Appl. January 2019:43-87. doi:10.1016/B978-0-12-814470-1.00003-4

[12] Sharma GN, Dave R, Sanadya J, Sharma P, Sharma KK. VARIOUS TYPES AND MANAGEMENT OF BREAST CANCER: AN OVERVIEW. J Adv Pharm Technol Res. 2010;1(2):109. /pmc/articles/PMC3255438/. Accessed November 10, 2021.

[13] Tang J, Ahmad A, Sarkar FH. The Role of MicroRNAs in Breast Cancer Migration, Invasion and Metastasis. Int J Mol Sci 2012, Vol 13, Pages 13414-13437. 2012;13(10):13414-13437. doi:10.3390/IJMS131013414

[14] Genome-wide association studies (GWAS). Genome.gov. (n.d.). Retrieved November 12, 2021, from https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies.

[15] Lee, Sangseon, et al. "Cancer Subtype Classification and Modeling by Pathway Attention and Propagation." Bioinformatics, vol. 36, no. 12, 2020, pp. 3818–3824., doi:10.1093/bioinformatics/btaa203.

[16] Wirapati P, Sotiriou C, Kunkel S, et al. Open Access Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. 2008. doi:10.1186/bcr2124

[17] Li X, Truong B, Xu T, Liu L, Li J, Le TD. Automated image analysis system for studying cardiotoxicity in human pluripotent stem cell-Derived cardiomyocytes. 2020. doi:10.1186/s12859-021-04215-3

[18] Matamala N, Vargas MT, González-Cámpora R, et al. Tumor MicroRNA expression profiling identifies circulating MicroRNAs for early breast cancer detection. Clin Chem. 2015;61(8):1098-1106. doi:10.1373/clinchem.2015.238691

[19] Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G.U. et al. Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. BMC Bioinformatics 20, 480 (2019). https://doi.org/10.1186/s12859-019-3050-8

[20] Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G.U. et al. Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. BMC Bioinformatics 20, 480 (2019). https://doi.org/10.1186/s12859-019-3050-8

[21] Yersal, Ozlem. "Biological Subtypes of Breast Cancer: Prognostic and Therapeutic Implications." World Journal of Clinical Oncology, vol. 5, no. 3, 2014, p. 412., doi:10.5306/wjco.v5.i3.412.

[22] Al-thoubaity, Fatma Khinaifis. "Molecular Classification of Breast Cancer: A Retrospective Cohort Study." Annals of Medicine and Surgery, vol. 49, 2020, pp. 44–48., doi:10.1016/j.amsu.2019.11.021.

[23] Wu, Jiande, and Chindo Hicks. "Breast Cancer Type Classification Using Machine Learning." Journal of Personalized Medicine, vol. 11, no. 2, 2021, p. 61., doi:10.3390/jpm11020061.

[24] Rehman, Oneeb, et al. "Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach." Cancers, vol. 11, no. 3, 2019, p. 431., doi:10.3390/cancers11030431.

[25] Søkilde, Rolf, Helena Persson, Anna Ehinger, Anna Chiara Pirona, Mårten Fernö, Cecilia Hegardt, Christer Larsson et al. "Refinement of breast cancer molecular classification by miRNA expression profiles." BMC genomics 20 (2019): 1-12.

[26] Souza, Karen CB, Adriane F. Evangelista, Letícia F. Leal, Cristiano P. Souza, René A. Vieira, Rhafaela L. Causin, Ana Caroline Neuber et al. "Identification of cell-free circulating microRNAs for the detection of early breast cancer and molecular subtyping." Journal of Oncology 2019 (2019).

[27] Sarkar, Jnanendra Prasad, Indrajit Saha, Anasua Sarkar, and Ujjwal Maulik. "Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers." Computers in Biology and Medicine 131 (2021): 104244.

[28] Richard, Vinitha, Matthew G. Davey, Heidi Annuk, Nicola Miller, Róisín M. Dwyer, Aoife Lowery, and Michael J. Kerin. "MicroRNAs in molecular classification and pathogenesis of breast tumors." Cancers 13, no. 21 (2021): 5332.

[29] Sideris, Nikolaos, Paola Dama, Salih Bayraktar, Thomas Stiff, and Leandro Castellano. "LncRNAs in breast cancer: a link to future approaches." Cancer Gene Therapy 29, no. 12 (2022): 1866-1877.

[30] Andreini, Paolo, Simone Bonechi, Monica Bianchini, and Filippo Geraci. "MicroRNA signature for interpretable breast cancer classification with subtype clue." Journal of Computational Mathematics and Data Science 3 (2022): 100042.

[31] Contreras-Rodríguez, Jorge Alberto, Diana Margarita Córdova-Esparza, María Zenaida Saavedra-Leos, and Macrina Beatriz Silva-Cázares. "Machine Learning and miRNAs as Potential Biomarkers of Breast Cancer: A Systematic Review of Classification Methods." Applied Sciences 13, no. 14 (2023): 8257.