PROTOTYPE OF DECISION SUPPORT SYSTEM USING PATTERN RECOGNITION AS AN APPLICATION OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR EARLY DIAGNOSIS OF GENETIC DISEASES

[1]PADMAVATHI P, [2]ANAGHA S SETLUR, [3]ADITHYA SABHAPATHI C, [4]SATYAM SURESH RAIKER, [5]SATYAM SINGH, [6]CHANDRASHEKAR K, [7]VIDYA NIRANJAN

[1]Department of Biotechnology, R V College of Engineering, Bengaluru – 560059
E-mail: vidya.n@rvce.du.in

Abstract -
With enormous amount of information existing for several diseases, managing and accessing the required data becomes tedious. Moreover, in the healthcare systems, the advent of technology has now proved to be extremely beneficial for storing and retrieving such data. An even more challenging task is the appropriate use of this data that can aid the clinicians, researchers and diagnosticians make an early decision that can help in formulating an appropriate treatment strategy for the patients. Thus, with this background, the present study aimed to design and develop a decision support system using pattern recognition and implementing artificial intelligence (AI) and machine learning (ML) algorithms. As a test case for constructing the model, polymorphisms from twenty cancer exome datasets were utilized for building the model. A basic pattern was recognized for 4181 cancer polymorphisms that were previously identified among various cancer exome datasets. All common nucleotide base substitutions that occurred most frequently were identified. After ruling out other machine learning algorithms in an initial scrutiny to identify the best working algorithm for DSS, a model was designed using supervised random forest technique by selecting relevant features. When polymorphisms from all twenty exome datasets were used for training, prediction for various cancer types were found to be correct. The accuracy of the designed model was found to be 79%. Thus, although the method proposed in the present study provides insights into early detection of various cancer types, the protocol can be applied for early diagnosis of other genetic diseases as well, thereby providing immense support for easy decision-making in healthcare and medicine.

Keywords - Decision Support System, Polymorphisms, Machine Learning, Random Forest, Disease Diagnosis

I. INTRODUCTION

With an enormous amount of information amassed for several diseases, it is essential to have an all-in-one platform that can both provide information and at the same time aid clinicians, diagnosticians and researchers take a decision on the type of treatment strategy by providing an early disease diagnosis. This can be achieved by designing and developing a clinical decision support system (DSS) that primarily improves healthcare systems by augmenting rates of faster medical diagnosis [1]. Traditionally, a DSS system is composed of software that is intended to be a direct aid to clinical decision making. Currently, it is being utilized more for early detection of disease probability so the patient can undergo appropriate treatment at the earliest [2, 3]. With advancements in technology, computer aided decision support systems often can be easily accessed, manipulated and monitored. A decision support system can be categorized into various types. Model-driven DSS models offer support for the decisions with the utilization of other analytical tools, knowledge-based systems provide a set of recommendations to solve the problem at hand by accessing the existing knowledge and data-driven systems retrieves the data, allows management and data manipulation [4, 5]. Several approaches have been used for designing a DSS including structural modelling and empirical assessment methods [6, 7]. However, with improvements in technology, artificial intelligence, machine learning and neural networks are now preferred [8, 9].

Machine Learning (ML) models enable the computer systems to acquire and learn from the data via statistical approaches [10]. Additionally, DSS models that use ML can be trained using the available medical data to corroborate the decision of a diagnostician/clinician [11]. Previous studies [12, 13, 14] have developed models by ML algorithms that diagnose patients with rare cancers, pulmonary diseases and neuromuscular diseases. Thus, DSS models are considered as clinical prototypes where data mining, integration, interpretation and access are available for easy and quick use. As an application of DSS, the present study focused on developing a prototype DSS for diagnosing genetic diseases.

Genetic diseases are caused due to alterations and abnormalities in the genes, which modify the function of the proteins. They also collectively affect a major portion of the world's population. Currently, there is limited information available that can aid clinicians differentiate between similar such genetic anomalies [15]. This causes issues in treating the patients and providing genetic counselling [15]. Thus, addressing such genetic diseases by early diagnosis, predicting the risk of developing genetic diseases and its treatment are important aspects for research. Some examples of precarious genetic diseases include sickle cell anaemia, cystic fibrosis and heart diseases.

Cancer is one such genetic disease that is challenging to treat at later stages. As the existing cancer scenario in the world is on a continuous ascent, overcoming the difficulties of incorrect treatment decisions and prognosis is the prime focus of research. Thus, in this regard, the present study focused on designing a DSS system that can aid in the early diagnosis of genetic diseases. Cancer was taken as the test case for developing the model using AI and ML.

II. MATERIALS AND METHODS

2.1 Pattern recognition

A basic pattern was recognized for all the polymorphisms that were identified in our previous study [16] from twenty cancer exome datasets. The .csv file containing the identified variants from our previous study contained all the variants and information related to every base modification that occurred in every chromosome. Using basic MS excel functions, all those base substitutions that occurred maximum number of times was categorised as common mutations. Those that occurred only once or twice were noted as unique alterations. The base substitutions were counted with respect to the chromosome numbers. The substitutions were observed with reference to the human genome. The recognised pattern was then utilized for the DSS model.

2.2 Model building using ML algorithms

To build the DSS model, supervised machine learning algorithms were utilized, as precision and accuracy were the major requirements [17]. Since the training of the data was done using labelled data, supervised machine learning algorithms were considered for the study. Four major supervised ML algorithms such as Naïve Bayes, Support Vector Machine, Random Forest and KNN were initially used to explore which ML method was the best for building a preliminary DSS model. Studies have shown that Naïve Bayes is a good algorithm that can be used for classification purposes [18]. Furthermore, KNN algorithm is a non-parametric approach known to predict the target classes in a simplified manner [19]. Support Vector Machine is as a robust classification tool that has been previously used for cancer genomic studies [20]. Additionally, RF approaches have also been used to classify the various feature vectors for Alzheimer's, a genetic disease [21]. Due to these reasons, since the present study also focused on the early diagnosis of genetic diseases, these four ML supervised algorithms were utilised. A data pre-processing was carried out prior to data training to evaluate the not-a-number (NaN) values and to convert all the categorical values into numerical values. An initial set of polymorphism features from the .csv variant file were selected to comprehend the best performing algorithm. The performance metrics obtained after executing each algorithm was noted. The best performing algorithm was selected for further analysis by training all 20 cancer exome datasets.

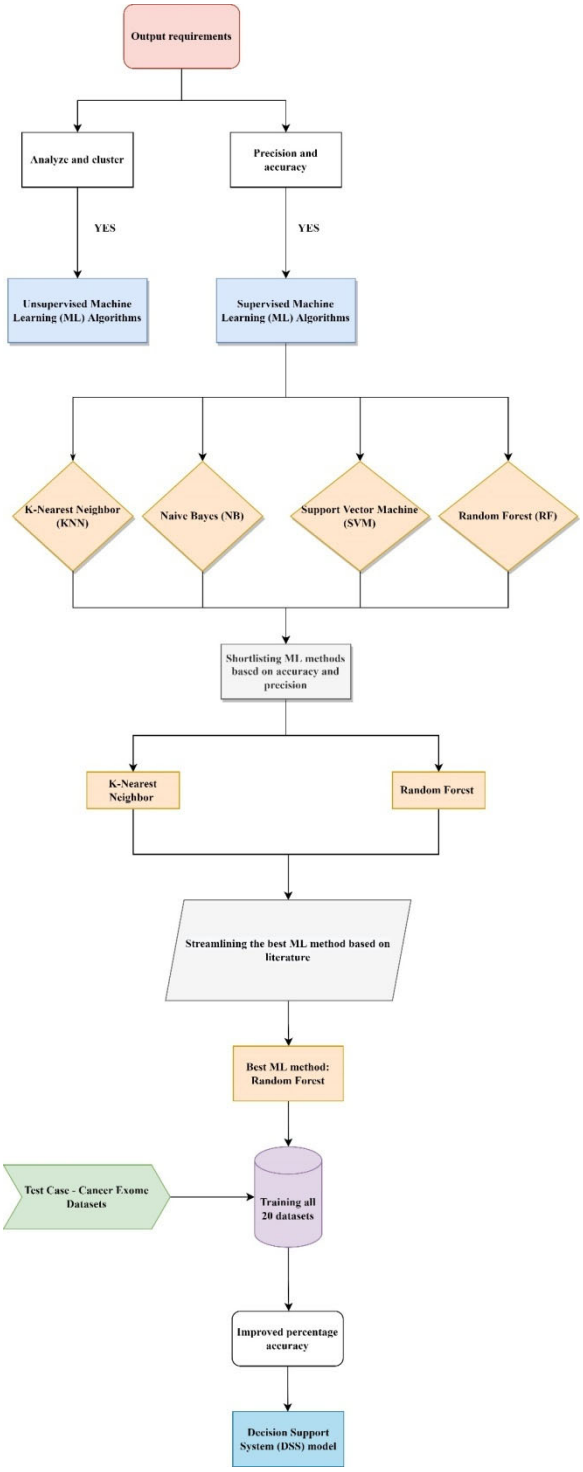


Fig. 1: Schematic representation of protocol for building DSS model for genetic diseases

2.3 Model improvement

In this method, the total number of features were increased to improve the model accuracy and precision. The features that were considered included the variant class, log odds that the variant is present in the tumor sample relative to the expected noise,

score for sorting the variants from tolerant to intolerant, allelic frequency of the sample, the type of variant after SIFT sorting, median base quality of each allele, median fragment length of each allele, median mapping quality of each allele, allelic depth of the sample, forward and reverse read counts for each allele, read depth, phred-scaled posterior probability that the alternate alleles are not germline variants, median distance from the end of the read for each alternate allele, population allele frequency of the alternate alleles and approximate read depth of the sample (https://support.sentieon.com/appnotes/out_fields/).

In each case, the ML algorithms were executed in Python using the Jupyter Notebook. The percentage accuracy was noted and the number of cancer types for which the predictions were accurate was also analysed. The entire protocol for arriving at the DSS model is illustrated in **Fig. 1.**

III. RESULTS AND DISCUSSION

3.1 Pattern recognition

In the initial .csv file that was obtained as an outcome from our previous study, 4181 SNPs (Single Nucleotide Polymorphisms) were identified. In the present study, for pattern recognition, common SNPs across all twenty exome datasets that revealed 4181 variations were identified. It was observed that base substitution from C to T and from G to A were found to be most common among all datasets.Further, the base alteration from G to A occurred in 16 out of 20 cancer exome datasets, while modification from C to T occurred in 17 out of 20 datasets. The base alterations from G to A appeared 191 timesout of 4181 and C to T appeared 198 timesin human diffuse type gastric cancer, highlighted in green in the table**(Table 1).** This recognized pattern helped in understanding important SNPs in cancer exomes that could be further used for detection of important and essential polymorphisms for early cancer detection.

Cancer Exome Dataset IDs taken as test case for model building	Cancer Type	Most Common SNPs and their frequency of occurrence in each dataset
SRR900123	Intrahepatic cholangiocarcinoma	G-A: 129 C-T: 127
SRR900099	Intrahepatic cholangiocarcinoma	G-A: 13 C-T: 8
SRR941052	Human Diffuse Type Gastric Cancer	C-T: 45 G-A: 37
SRR941053	Human Diffuse Type Gastric Cancer	C-T: 28 G-A: 21
SRR941054	Human Diffuse Type Gastric Cancer	C-T: 198 G-A: 191
ERR166304	Non BRCA1/BRCA2 familial breast cancer	A-C: 39 G-T: 31 C-A: 31
ERR166307	Non BRCA1/BRCA2 familial breast cancer	G-T: 14 C-A: 11 C-T: 11 A-G: 11
ERR166310	Non BRCA1/BRCA2 familial breast cancer	C-T: 13 G-A: 10
ERR166312	Non BRCA1/BRCA2 familial breast cancer	C-T: 18 G-A: 14
ERR166335	Non BRCA1/BRCA2 familial breast cancer	C-T: 10 G-A: 8
ERR232255	Pancreatic adenocarcinoma	G-A: 35 C-T: 25
ERR232254	Pancreatic adenocarcinoma	G-A: 31 C-T: 29
ERR232253	Pancreatic adenocarcinoma	G-A: 33 C-T: 26
ERR035489	High-grade serous ovarian cancer	C-T: 39 G-A: 33
ERR035488	High-grade serous ovarian cancer	C-T: 38 G-A: 36
ERR035487	High-grade serous ovarian cancer	G-A: 36 C-T: 36
ERR166336	Non BRCA1/BRCA2 familial breast cancer	G-T: 12 G-A: 12
ERR166303	Non BRCA1/BRCA2 familial breast cancer	C-T: 14 G-T: 11
SRR894452	Intrahepatic cholangiocarcinoma	A-T: 61 T-A: 55
SRR941051	Human Diffuse Type Gastric Cancer	C-T: 57 G-A: 46

Table 1: Pattern recognition for cancer exome datasets as test case for building DSS model

Previous studies have stressed on the importance of using pattern recognition for identification of varied cancer types and states that to better predict occurrence of genetic diseases such as cancer, integrative patterns that result from omics data analysis such as genomics, transcriptomics, proteomics and metabolomics, contribute to precision medicine [22]. The present study has also identified patterns from polymorphism data that were detected using omics approaches in our previous study [16]. Furthermore, computer-aided recognition of genetic syndromes has been previously carried out using intelligent phenotypic patterns [23]. Moreover, pattern recognition has been previously used to aid diagnosis of lung cancer using deep learning algorithms [24]. The present study however, has recognised certain basic patterns that aid in easy cancer detection and which can be incorporated in similar DSS models. Typically, recurrent base substitutions from C to Thas been previously identified in metastatic melanomas [25], and high rates of base changes from G to A has been observed in gall bladder carcinomas [26]. Thus, the present study identified valuable patterns that can point towards early disease diagnosis.

3.2 DSS model and its improvement

Initially, when four supervised ML algorithms were used, the percentage of accuracy was not very high. When KNN supervised ML algorithm was used, the

weighted average for precision was found to be 0.40, 0.41 for recall and 0.40 for F1-score. The accuracy of the model was found to be 41%. Likewise, the weighted average score for precision was found to be 0.11, 0.34 for recall and 0.17 for F1-score when SVM model was employed. The accuracy of the model was observed to be 34%. For Naïve Bayes, the weighted average for precision was found to be 0.24, 0.34 for recall and 0.19 for f1-score. The accuracy for NB model was 32%. With RF algorithm, the weighted average was observed to be 0.39 for precision, 0.41 for recall and 0.40 for F1-score. Model accuracy when RF was used was observed to be 41%**(Fig. 2).** Thus, although KNN,and RF models showed a higher accuracy, it was still lesser than expected and the output predictions were incorrect due to lesser training scores. Hence, the model was improved by selecting additional features and further streamlining our algorithms to RF and decision tree, all twenty exome datasets were trained.

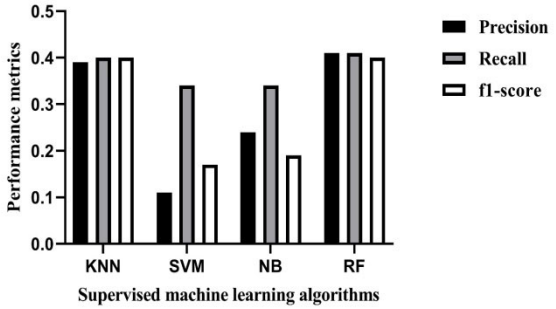


Fig. 2: Initial performance metrics showing precision, recall and f1-score values for four supervised machine learning algorithms K-nearest neighbour, Support vector machine, Naïve Bayes and Random Forest

When a greater number of features were chosen for training the datasets, for random forest model, the weighted average for precision, recall and f1-score was found to be 0.79. Although both SVM and RF showed similar accuracy in the preliminary trials, only RF algorithm was taken forward for model improvement since literature suggests that RF model is more accurate, robust and has a better overall performance thanKNN [27].The model accuracy for RFwas observed to be 79%, an increase in the accuracy by 38%. A detailed analysis of the accuracies for all models used in the paper is provided in **Fig. 3.** Additionally, the predictions for all five cancer types that were taken as test case for the present study were also found to be correct. These cancer types included high-grade serous ovarian cancer, pancreatic adenocarcinoma, non BRCA1/BRCA2 familial breast cancer, human diffuse-type gastric cancer and intrahepatic cholangiocarcinoma. Therefore, the present DSS prototype serves as the basic model for developing similar such models for other genetic diseases.

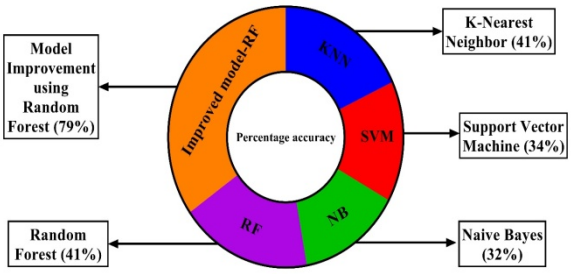


Fig. 3: Categorization of various ML algorithms showing the percentage accuracies of the models when DSS was built using polymorphisms from cancer exome datasets as a case study

Previously, a fuzzy decision support system using genetic algorithm has been designed for early diagnosis of heart disease [28]. The model also predicts the level of risk associated with genetically caused heart problems. Likewise, another study developed a machine learning based DSS for heart disease diagnosis, wherein the study utilized genetic algorithms, fuzzy clustering, random forest, independent component analysis and bayes theorem for predictions [8, 29, 30]. Another study used SVM and artificial neural networks for developing a DSS system that predicts presence or absence of heart disease, wherein the SVM model showed an accuracy of 80.41% [31]. Thus, the present study designed a model DSS system using polymorphisms from cancer datasets as the major inputs as a test case, which revealed RF algorithm as a fine approach to further develop similar such models for other related genetic diseases.

IV. CONCLUSIONS AND FUTURE PERSPECTIVES

With the rise of several genetic diseases, the need for detection of these diseases is the primary concern. Advent of omics technology has now enabled researchers to access and generate huge amounts of data that can be valuable for identification of several disease-causing polymorphisms. As detecting the different types of similar genetic diseases is difficult, comprehending the various disease-causing polymorphisms and identifying essential patterns is vital for early disease diagnosis. Using computer aided AI and ML techniques accelerates and simplifies many issues during diagnosis. Therefore, taking a decision on the diagnosis and treatment strategies becomes easier for the patients. Hence, the current study focused on using AI and ML techniques to design and develop a model decision support system that can aid in the predicting the probability of occurrence of genetic diseases, taking cancer as the test case for building the model. Initially, the four different supervised ML algorithms such as KNN, Naïve Bayes, SVM and RF were utilized, among which only KNN and RF showed a better accuracy compared to the other models. Since RF method is typically considered as the better of two, RF was then

utilized for model improvement by employing more features, which resulted in an accuracy of 79%, a jump from our initial model with 41% accuracy. Thus, our model serves as a prototype for developing similar such models for other genetic diseases, and paves the way for early disease diagnosis. As future perspective, this research work will be continued further to make additional improvements to the devised model.

CONFLICTS OF INTEREST

All the authors declare no conflict of interest and consent to the publication of the article.

ACKNOWLEDGEMENTS

We would like to thank all the staff and non-staff members of RV College of Engineering for providing the infrastructural support for completing this research work. Special thanks to Mr. Akshay Uttarkar for reviewing the manuscript.

REFERENCES

[1] Osheroff, J. A., Teich, J. M., Levick, D., Saldana, L., Velasco, F. T., Sittig, D. F., ... & Jenders, R. A. (2012). *Improving outcomes with clinical decision support: an implementer's guide*. Himss Publishing.

[2] Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., & Tang, P. C. (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, *8*(6), 527-534.

[3] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, *3*(1), 1-10.

[4] Moreira, M. W., Rodrigues, J. J., Korotaev, V., Al-Muhtadi, J., & Kumar, N. (2019). A comprehensive review on smart decision support systems for health care. *IEEE Systems Journal*, *13*(3), 3536-3545.

[5] Holsapple, C. W. (2008). DSS architecture and types. In *Handbook on Decision Support Systems 1* (pp. 163-189). Springer, Berlin, Heidelberg.

[6] Ramamurthy, K., King, W. R., & Premkumar, G. (1992). User characteristics—DSS effectiveness linkage: An empirical assessment. *International Journal of Man-Machine Studies*, *36*(3), 469-505.

[7] Alshibly, H. H. (2015). Investigating decision support system (DSS) success: A partial least squares structural equation modeling approach. *Journal of Business Studies Quarterly*, *6*(4), 56.

[8] Safdar, S., Zafar, S., Zafar, N., & Khan, N. F. (2018). Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. *Artificial Intelligence Review*, *50*(4), 597-623.

[9] Mohammed, M. A., Abd Ghani, M. K., Arunkumar, N. A., Hamed, R. I., Mostafa, S. A., Abdullah, M. K., & Burhanuddin, M. A. (2020). Decision support system for nasopharyngeal carcinoma discrimination from endoscopic images using artificial neural network. *The Journal of Supercomputing*, *76*(2), 1086-1104.

[10] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920-1930.

[11] Fraccaro, P., Plastiras, P., Dentone, C., Di Biagio, A., & Weller, P. (2015). Behind the screens: Clinical decision support methodologies–A review. *Health Policy and Technology*, *4*(1), 29-38.

[12] Rother, A. K., Schwerk, N., Brinkmann, F., Klawonn, F., Lechner, W., & Grigull, L. (2015). Diagnostic support for selected paediatric pulmonary diseases using answer-pattern recognition in questionnaires based on combined data mining applications—a monocentric observational pilot study. *PLoS One*, *10*(8), e0135180.

[13] Grigull, L., Lechner, W., Petri, S., Kollewe, K., Dengler, R., Mehmecke, S., ... & Klawonn, F. (2016). Diagnostic support for selected neuromuscular diseases using answer-pattern recognition and data mining techniques: a proof-of-concept multicenter prospective trial. *BMC medical informatics and decision making*, *16*(1), 1-9.

[14] Sidiropoulos, K., Glotsos, D., Kostopoulos, S., Ravazoula, P., Kalatzis, I., Cavouras, D., & Stonham, J. (2012). Real time decision support system for diagnosis of rare cancers, trained in parallel, on a graphics processing unit. *Computers in biology and medicine*, *42*(4), 376-386.

[15] Pogue, R. E., Cavalcanti, D. P., Shanker, S., Andrade, R. V., Aguiar, L. R., de Carvalho, J. L., & Costa, F. F. (2018). Rare genetic diseases: update on diagnosis, treatment and online resources. *Drug discovery today*, *23*(1), 187-195.

[16] Padmavathi, P., Setlur, A. S., Chandrashekar, K., & Niranjan, V. (2021). A comprehensive in-silico computational analysis of twenty cancer exome datasets and identification of associated somatic variants reveals potential molecular markers for detection of varied cancer types. *Informatics in Medicine Unlocked*, 100762.

[17] Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Academic pathology*, *6*, 2374289519873088.

[18] Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88-91.

[19] Rajaguru, H. (2019). Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. *Asian Pacific journal of cancer prevention: APJCP*, *20*(12), 3777.

[20] 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 genomics & proteomics*, *15*(1), 41-51.

[21] Xu, L., Liang, G., Liao, C., Chen, G. D., & Chang, C. C. (2019). K-skip-n-gram-RF: a random Forest based method for Alzheimer's disease protein identification. *Frontiers in genetics*, *10*, 33.

[22] Cheng, T., & Zhan, X. (2017). Pattern recognition for predictive, preventive, and personalized medicine in cancer. *EPMA Journal*, *8*(1), 51-60.

[23] Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., ... & Gripp, K. W. (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine*, *25*(1), 60-64.

[24] Sun, W., Zheng, B., & Qian, W. (2016, March). Computer aided lung cancer diagnosis with deep learning algorithms. In *Medical imaging 2016: computer-aided diagnosis* (Vol. 9785, p. 97850Z). International Society for Optics and Photonics.

[25] Wong, S. Q., Behren, A., Mar, V. J., Woods, K., Li, J., Martin, C., ... & McArthur, G. A. (2015). Whole exome sequencing identifies a recurrent RQCD1 P131L mutation in cutaneous melanoma. *Oncotarget*, *6*(2), 1115.

[26] Li, M., Zhang, Z., Li, X., Ye, J., Wu, X., Tan, Z., ... & Liu, Y. (2014). Whole-exome and targeted gene sequencing of gallbladder carcinoma identifies recurrent mutations in the ErbB pathway. *Nature genetics*, *46*(8), 872-876.

[27] Lind, A. P., & Anderson, P. C. (2019). Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLoS One*, *14*(7), e0219774.

[28] Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. (2016, May). Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In *2016 5th*

International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 145-150). IEEE.

[29] Björk, J., Ekelund, U., & Ohlsson, M. (2012). Risk predictions for individual patients from logistic regression were visualized with bar–line charts. *Journal of clinical epidemiology*, *65*(3), 335-342.

[30] Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014). A machine learning system to improve heart failure patient assistance. *IEEE journal of biomedical and health informatics*, *18*(6), 1750-1756.

[31] Gudadhe, M., Wankhade, K., & Dongre, S. (2010, September). Decision support system for heart disease based on support vector machine and artificial neural network. In *2010 International Conference on Computer and Communication Technology (ICCCT)* (pp. 741-745). IEEE.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
DSS	Decision Support System
NB	Naïve Bayes
SVM	Support Vector Machine
RF	Random Forest
KNN	K-nearest neighbor
SNP	Single Nucleotide Polymorphism
A	Adenine
T	Thymine
G	Guanine
C	Cytosine

★★★