Precision Medicine: Cloud Pipeline for Breast Cancer Diagnosis and Treatment Recommendation

Lei Jiang, Deepti Kunupudi, and Michael J. Wolfe

Abstract— Breast cancer is one of the most common diseases in the world and diagnosing at the initial stages is critical for a positive long term prognosis. There are enormous amounts of genetic data generated from genomic tests due to the lower cost of Next Generation Sequencing (NGS) and the need for precision medicine. Whole Exome Sequencing (WES) data for patients can easily reach 10GB-15GB per patient. Thus, the size of the database can easily reach terabytes, with only 100 patients or more. Variant analysis or deploying machine learning algorithms on WES data is necessary for precision medicine, and usually terabytes of data are needed for an accurate result. It takes a supercomputer to perform the analysis on-premises; however, such technology is not always available to all research institutes or companies. Cloud computing solutions such as Google Cloud Platform (GCP), Amazon Web Services (AWS), and Microsoft Azure provide an ideal solution for storage and computation of these types of databases. In this paper, we will evaluate an on-premises breast cancer screening application and migrate it to Google Cloud Platform (GCP) with more enhanced data and user security. As a part of the solution, we will provide predictive analytics capabilities in the cloud.

Index Terms—Bioinformatics Databases, Cloud migration, Cloud Computing, Cancer Diagnosis Prediction, Genomic testing, Machine Learning, Precision Medicine, Neural Networks, GCP.

I. Introduction

Cloud computing has, in recent years, revolutionized computing and storage. Many companies see cloud computing as an opportunity to gain a competitive edge and meet business objectives. Additional cost flexibility, which favors pay per use models rather than upfront purchase, has attracted many entities ranging from large enterprises to small startups to academic institutions to establish cloud platforms. Apart from this, the scalability, reliability, and interoperability of cloud environments are also attractive features from which enterprises of all stripes can benefit. [20]. The range of cloud-based services offered grows simultaneously with the emergence of varying cloud service providers. [21] The most prominent leaders in the space are Google, Amazon, and Microsoft.

The motivation of migrating breast cancer and genome applications to the cloud is to reduce the costs of storing and computing an ever-growing dataset. Volumes of genomics and proteomics data increase exponentially. As the cost of genome sequencing drops from 10 million dollars a decade ago to 1 thousand dollars today, an enormous amount of data generated from genomic tests emerged as the market expanded. For example, the size of the genomics database Sequence Read Archive (SRA) grows from 1 terabyte in 2009 to more than 10,000 terabytes in 2019. That is 10,000 times growth in the last decade.

Additionally, raw sequencing data in public archives are doubling in size every 18 months [17]. Apart from cost and compute power, the decision is primarily based on business strategy, cloud environment, service suitability, risk assessment, vendor evaluation, and implementation we discussed before. This paper provides the framework and model for application migration to the cloud and leveraging cloud computing and storage for building precision medicine machine learning models for the purpose of detecting breast cancer.

II. Related Works

A. The rise of precision medicine

Precision medicine is defined as selecting the most effective treatments based on specific biomarkers in a patient's tumor and genetic sequence. Genomic testing identifies the patient's gene expression profiles to determine the corresponding sensitive targeted therapies. Precision medicine delivers individually tailored therapy based on the patient's disease subtype [2]. This approach ensures patients receive a treatment from which they would benefit most, avoid unnecessary treatments, reduce toxicity, and deliver improved outcomes.

Multiple studies have demonstrated the benefit of precision medicine [9, 13, 14]. For example, the Oncotype DX for Breast cancer, Colon cancer, and Prostate cancer is now a gold standard clinical genomic test that has delivered much better outcomes for those diseases, with over 1 million patients tested across 90 countries. Information from these genomic tests can help both patients and providers decide on treatment methods such as Chemotherapy, Radiation, aggressive treatment, or surgery. Other tests include EGFR mutations, ALK rearrangements, ROS1 fusions, and PD-L1 expression testing, which are recommended for advanced non–small cell lung cancer. These tests help with decisions on immunotherapies. Excitingly, research published this month suggested the

potential genomic test for Pancreatic Cancer by identifying Biomarkers in Pancreatic Cancer patients [11].

Precision medicine is not just limited to cancer diagnosis and treatment. It can also be applied to a variety of diseases such as diabetes and Alzheimer's, and prevention of diseases from actionable insights gained from data analysis. Therefore, precision medicine depends largely on rationalizing diverse sources of data, such as clinical, genomics, transcriptomics, or even environmental factors such as diet style and exercise habit, etc. Precision medicine could potentially find a cure for many lethal diseases.

B. *Cloud Computing*

Cloud computing is not just about technological improvements to data centers; it represents a fundamental change in how IT is provisioned and used. Risk-benefit and general impact analysis is critical for enterprises to decide and shift to cloud computing [22].

There are two widely used cloud computing classifications:

1. Deployment– Public, Private, Hybrid
2. Capabilities – Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS)

The main advantages of hosting the system on the cloud are:

- Highly scalable
- More flexible
- Reduced infrastructure cost
- Higher security
- Backup and disaster recovery
- No location constraints

Apart from the advantages, there are few challenges with cloud implementation as well. The following aspects need to be addressed or looked at closer to evaluate options before we implement/migrate it to the cloud.

- Security and Privacy
- Environment limitations
- Multitenancy
- Licensing
- Compliance requirements
- Network performance

III. CLOUD MIGRATION

This section will describe the application migration system to the cloud. The new cloud architecture for the migrated application along with the added flexibility, improvements, and capabilities to the new architecture on the cloud.

A. *Preliminary Analysis*

Before commencing the migration, we conducted a careful analysis of the advantages and disadvantages of cloud computing, which are referred to in the above section.

B. *Current Implementation*

A breast cancer screening data warehouse application is sitting on a local machine / on-premises which uploads the data in the database manually whenever a file is available. It leverages Python scripts as an ETL, and predictive machine learning capabilities are built based on a script, and basic reporting is available through the front-end.

C. *Proposed Implementation*

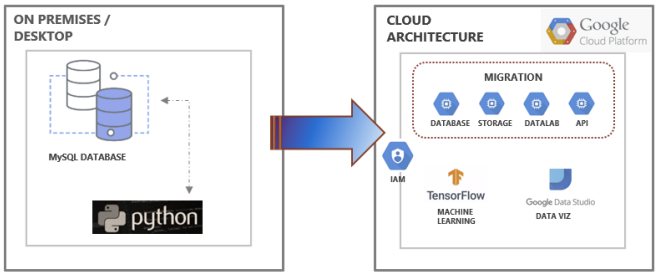


Fig.1. High-Level Proposed Architecture Diagram

The application will be migrated to the cloud and leverage cloud capabilities to provide and enhance data access and security. We will also provide more optimized integration, reporting capabilities, and augment computing to build more optimized machine learning algorithms and neural networks to detect breast cancer and provide precision medicine recommendations.

D. *Migration*

Applications from the local machine will be migrated to Google Cloud Platform using a sequential approach. As part of this approach, we have identified different phases through which we migrated the application(s).

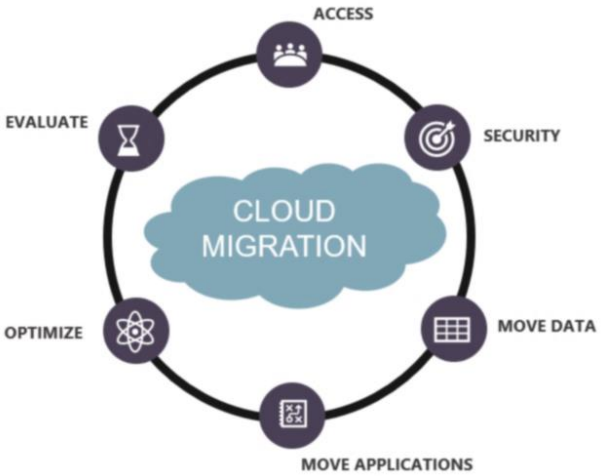


Fig.2. Strategical Migration Phases

The following are the high-level strategical phases -

1. Assess:

In this phase, we identified all the processes and components and mapped them to cloud pre-requisites based on the current application(s). We ensured that all the application(s) and dependent components are listed-down and checked the

suitability to the cloud platform, including hardware, performance, and application dependencies,

2. Setup Cloud Environment & Security

Moving the data off-premises is a strong first step in ensuring security, allowing the team to leverage Cloud IAM and authentication/authorization functions. The raw data was first loaded to Google Cloud Storage, which is only accessible to the project owners. A subnet was created to contain the data engineering and machine learning pipelines, which is only accessible via service accounts – this secures data on the fly. Finally, front end data access is only permitted via DataStudio, which will be limited only to authorized users.

3. Move data

There are many ways of migrating the data to the cloud, including batch transfers, database dumps, offline disk imports, or streaming to persistent disks alongside cloud storage options. For ease of migration, data was dumped into CSV files and migrated to Google Cloud Storage. Later the data was migrated to BigQuery via ETL processes with some optimizations added based on the data frequency and usage for different reports.

4. Move applications

With the data migrated, the applications and their dependencies were migrated to GCP. We first tested the models in DataLab, a self-contained Notebook server, which let us keep a close watch on the processes and confirm the applications and their dependencies are properly migrated and verified.

5. Optimize

Leveraging the cloud architecture and flexibility, new applications were deployed for integrating the data using Dataflow and Dataproc, as well as automating the data integration using Google Composer, which manages the workflow orchestration. Additionally, standing up a DataStudio server for reporting directly from database gives end-users much more flexibility in defining their metrics and visualizations.

E. Cloud Architecture

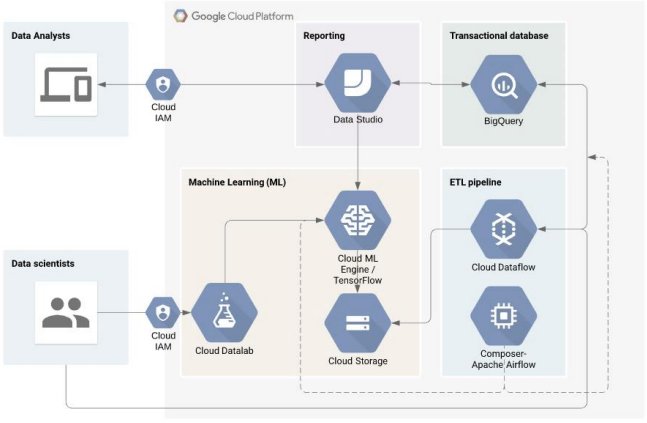


Fig.3. Cloud Architecture

F. Cloud Monitoring

Cloud Monitoring makes it easier to identify patterns and discover potential security risks in the infrastructure dashboard. Key capabilities are:

- Ability to monitor large volumes of data across many locations
- Gain visual footprint across the applications within the cloud.

IV. PREDICTIVE ANALYTICS

A. Background

Tools like DataLab and Dataproc make developing and deploying machine learning models on Google Cloud Platform (GCP) to facilitate classification to detect malign or benign tumors a much smoother process than on-premises solutions. Lab environments facilitate initial analysis that informs algorithm creation, along with measuring prediction accuracy of said algorithms and adjusting accordingly.

B. Building end to end ML pipeline:

The quality of preprocessing standardization is critical. Dataflow pipelines verify the data is clean and has no missing data. The end to end machine learning pipeline on GCP includes preprocessing, feature engineering, and classifications. The general flow:

Fig.4. ML Pipeline

Pre-Processing: In our machine learning model, we utilize preprocessing techniques to normalize and eliminate redundant and ambiguous data from the dataset. The breast cancer dataset consists of 31 variables and 569 observations. The dataset supports binary classification models since it has only two class labels: benign and malignant. [25] The dataset has no missing values. The next step is feature selection, which will improve the accuracy of the model.

Feature Selection: The system selects the best features from all variables to improve the performance and accuracy of the system. This ensures the best features are selected without overfitting the model.

C. Classification

Since this is a binary classification exercise, the performance is evaluated by using different models within the classification area. Three models – logistic regression, random forest, and neural networks will be deployed; we will then review the model performance and accuracy for each.

Logistic Regression:

The central mathematical concept that underlies logistic regression is the logit – the natural algorithm of an odds ratio. Logistic regression is well suited for describing and testing hypotheses about relationships between categorical outcomes and one or more categorical or continuous predictor variables. [26]

Below is the DataLab-facilitated data exploration for our models in the Google Cloud Platform.

Dataset info

Number of variables	30
Number of observations	569
Total Missing (%)	0.0%
Total size in memory	133.4 KiB
Average record size in memory	240.1 B

Variables types

Numeric	20
Categorical	0
Boolean	0
Date	0
Text (Unique)	0
Rejected	10
Unsupported	0

Fig 5. Dataset Information using Datalab

DataLab also provided an auto-exploratory data analysis to identify trends and variations. Given details about the correlation between the predictors are below, which were used to select the appropriate features for the models.

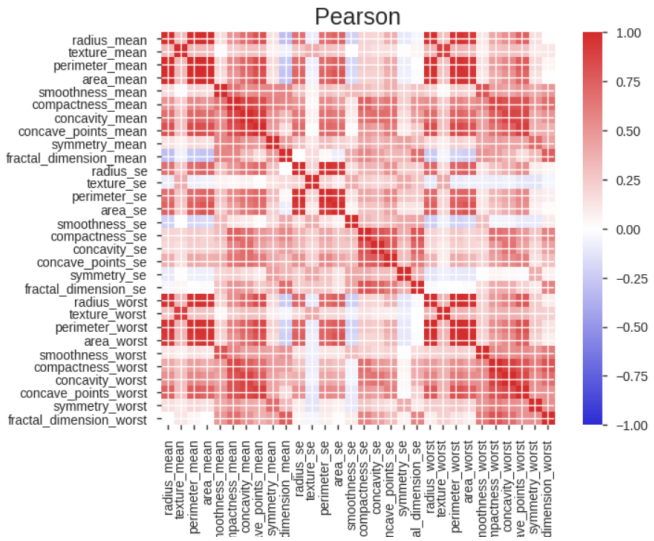


Fig 6. Correlation Matrix

After feature selection, the models were deployed using a standard train/test split. The following confusion matrix explains the results of the logistic regression model.

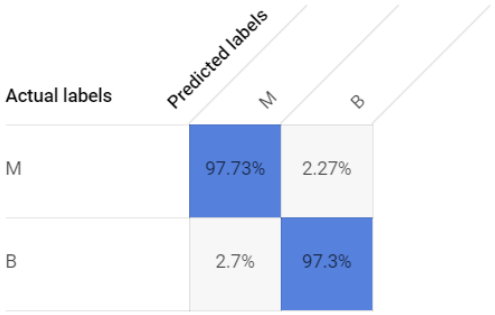
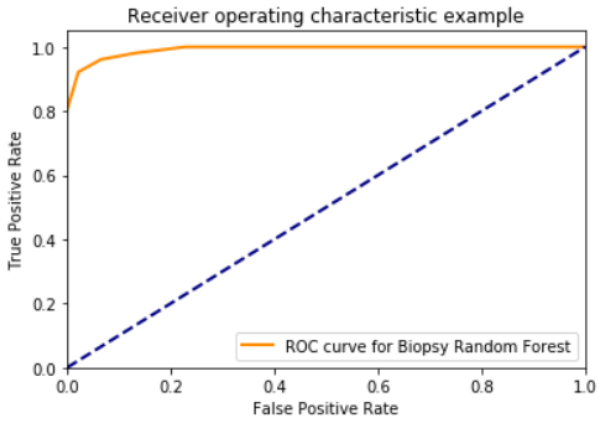


Fig.7 Confusion Matrix

The accuracy of the logistic regression is around 0.9746, a strong insight into the prediction accuracy of the simple initial model.



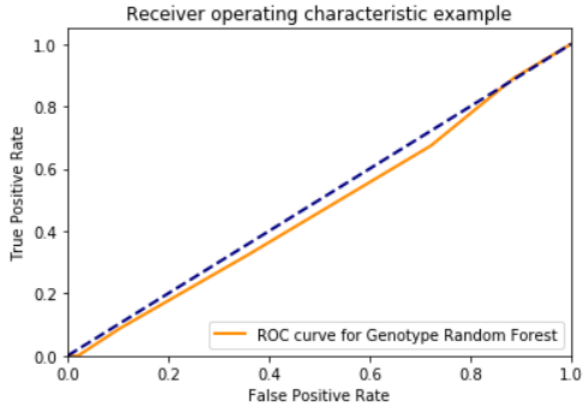


Fig.8 ROC Curve for Random Forest

Random Forest:

The random forest algorithm forms a family of classification methods that rely on iterations of several decision trees. This ensemble of classifiers is based on components that are grown from a certain amount of randomness. [27]

Our random forest was implemented using DataLab in the cloud. The accuracy of one random forest is around 0.9371. It provides similar metrics and insights in the prediction accuracy based on the simple initial model.

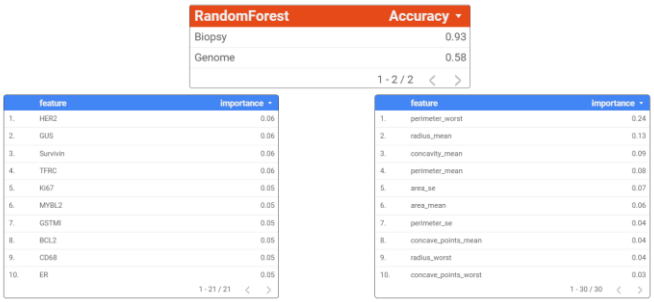


Fig.9 Feature Selection

Neural Networks:

Neural networks are a type of machine learning algorithm that can be used for regression and classification problems. For classification purposes, they are often used in comparison with logistic regression and discriminant analysis. Their advantages lie in the robustness and ability to work with missing data.[26]

The feature selection listed above acts as input for the neural network. The data is split 80:20, and the model is generated using TensorFlow. In this NN we are leveraging the default layer size with an [10,20,10] array, indicating that there will be a layer of 10 neurons which are connected to 20 neurons in the next layer and each of which is connected to 10 neurons in the next layer. The accuracy of the model is around 0.8132.

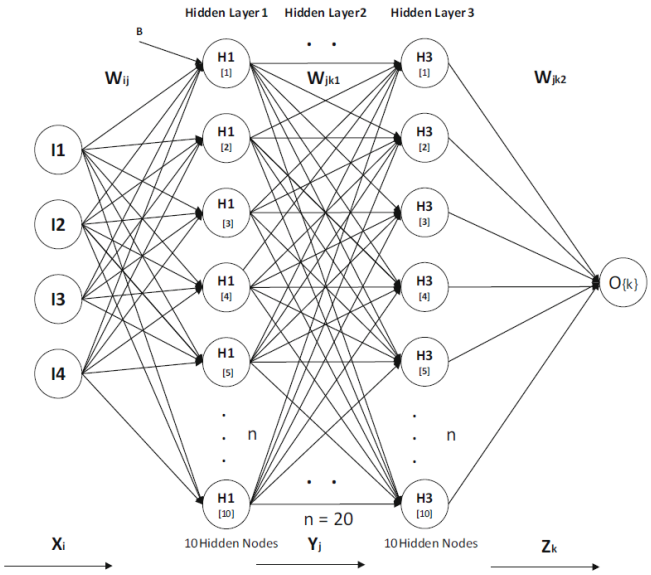


Fig 10. Neural Network Architecture [25]

Prediction Matrix

Model	Accuracy
Logistic Regression	0.9746
Random Forest	0.9371
Neural Network	0.8132

Fig.11 Prediction matrix across models

Conclusion

Each of the models provides different accuracies based on the classification of the model; however, the results provided illustrate the powerful Google Cloud computing capabilities in general. Even though the results look promising from a statistical perspective, additional research could translate to different options.

V. Conclusion and Future Work

This paper outlines the pre-work and understanding of cloud migration as well as high-level cloud migration strategy for moving data and applications to the cloud. It also outlines the creation of predictive models for breast cancer using logistic regression, random forest, and neural networks on biopsy and genome data. During the cloud pre-migration step, understanding the necessary steps required for cloud migration and detailing the need for the cloud are critical knowledge steps to ensure future success. Once a concrete understanding is established, forming a robust migration strategy for a smooth transition to the cloud from on-premises and build required pipelines for data integration and reporting serves to execute on the learnings gained from the first step. For predictive modeling, the first stage of modeling is to extract the feature selection required for input. We created logistic regression, random forest, and neural network models to discover the most accurate model for providing the highest classification rate on the validation set.[26]

The study performed is at high-level to evaluate the understanding of cloud migration and capabilities, which can be leveraged for building different models. Further research is needed for each of these models to add more nuance and improve accuracy. Even though the results provided based on the initial dataset of data looking promising, not all factors within the data are considered; taking these elements into account could engender further research opportunities.

Acknowledgment

This study was performed as a part of school case study on cloud migration and model evaluation using cloud solutions. The information is referenced from different articles related to cloud and machine learning models.

References

[1] Montague E, Stanberry L, Higdon R, et al. MOPED 2.5--an integrated multi-omics resource: multi-omics profiling expression database now includes transcriptomics data. OMICS. 2014;18(6):335-43.
[2] Servant N, Roméjon J, Gestraud P, et al. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. Front Genet. 2014;5:152. Published 2014 May 30. doi:10.3389/fgene.2014.00152
[3] Zhang, W., Li, F., & Nie, L. (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology, 156 Pt 2, 287-301

[4] Kim, Minseung & Rai, Navneet & Zorraquino, Violeta & Tagkopoulos, Ilias. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nature Communications. 7. 13090. 10.1038/ncomms13090.
[5] Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534(7605):55-62.
[6] Zhang H, Liu T, Zhang Z, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell. 2016;166(3):755-765.
[7] Mackey et al. Relational Databases for Biologists
[8] Janetzki et al. Genome Data Management using RDBMSs. Technical Report. 2015.
[9] McVeigh et al. Clinical use of the Oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. Breast Cancer - Targets and Therapy. 2017; 9:393-400
[10] Carlsson et al. HER2 expression in breast cancer primary tumours and corresponding metastases. Original data and literature review. Br J Cancer. 2004; 90(12): 2344–2348.
[11] Dimitrakopoulos et al. Identification and Validation of a Biomarker Signature in Patients With Resectable Pancreatic Cancer via Genome-Wide Screening for Functional Genetic Variants. JAMA Surg. 2019;3: e190484. doi: 10.1001/jamasurg.2019.0484.
[12] Jiang et al. Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies. Cancer Cell. 2019; 35: 428-440
[13]J Mascaux et al. Genomic Testing in Lung Cancer: Past, Present, and Future. Natl Compr Canc Netw. 2018 Mar;16(3):323-334. doi: 10.6004/jnccn.2017.7019.
[14] Katherine et al. Precision Oncology Decision Support: Current Approaches and Strategies for the Future. Clinical Cancer Research. 2018. DOI: 10.1158/1078-0432.CCR-17-2494
[15] van Dongen et al. Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. Blood. 2015. 125:3996-4009; DOI: https://doi.org/10.1182/blood-2015-03-580027
[16] https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data
[17] Langmead et al. Could computing for genomic data analysis and collaboration. Nature Reviews Genetics. 2018. 19, 208-219
[18] https://opencirrus.org/benefits-cloud-computing/
[19] Tucci et al. Design and evaluation of a genomics variant analysis pipeline using GATK Spark tools. Computer Science department of Cornell University. 2018.
[20] Cloud Migration Research: A Systematic Review Pooyan Jamshidi, Aakash Ahmad, and Claus Pahl
[21] Migration to Cloud Computing: A Decision Process Model - Adel Alkhalil, Reza Sahandi, David John
[22] Cloud Migration: A Case Study of Migrating an Enterprise IT System to IaaS - Ali Khajeh-Hosseini David Greenwood Ian Sommerville
[23] Migration of an on-premise application to the Cloud: Experience report - Pavel Rabetski and Gerardo Schneider [24] Steps to the cloud - Clarnet
[25] Breast Cancer Classification Using Deep Neural Networks S. Karthik, R. Srinivasa Perumal and P. V. S. S. R. Chandra Mouli [26] Predicting company growth using logistic regression and neural networks - Marijana Zekić-Sušac1,†, Nataša Šarlija1, Adela Has1 and Ana Bilandžić1 [27] Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic Cuong Nguyen, Yong Wang, Ha Nam Nguyen