A SURVEY ON CANCER SUBTYPING BASED ON DEEP LEARNING USING PAN-CANCER AND MULTIOMIC DATA

S.Keerthana[1], Ms.K. Shanmugapriya.M.E[2]
[1]PG Scholar, Department of Computer Science and Engineering, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science and Engineering, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India

ABSTRACT

Tumor subclasses with clinical implications are identified by breast cancer gene expression patterns. In this study, ROBERT TIBSHIRANI et al. make a suggestion. Based on changes in gene expression, the tumours were divided into three groups: basal epithelial-like, ERBB2-overexpressing, and normal breast-like. Two separate gene sets, one representing a collection of 456 cDNA clones originally chosen to reflect intrinsic tumour features and the other being a gene set that was highly correlated with patient prognosis, were clustered to reveal that both groupings were quite robust. The basal-like subtype had a dismal prognosis, and the two oestrogen receptor-positive groups had significantly different outcomes, according to survival studies on a sub cohort of patients. In this study, three fibroadenomas and 78 breast carcinomas were examined. This collection comprises of 40 tumours that have already been studied and characterised. A total of 85 tissue samples from 84 individuals were examined.

Keywords: Deep learning, genomics, multi-omics, semi supervised learning, variational autoencoder

1. INTRODUCTION

Molecular stratification has shown to be a useful tool for predicting the clinical outcome of BRCA patients and assisting doctors in therapeutic decision-making throughout time. BRCA is known to be a diverse and fatal malignant illness. The so-called intrinsic BRCA molecular subtypes Luminal A, Luminal B, Her2-enriched, Basal, and Normal-like, which were first identified in the early 2000s using unsupervised hierarchical clustering on BRCA microarray gene expression patterns, have steadily grown in importance as prognostic indicators in clinical practise. In recent years, a number of multi-gene prognostic molecular tests have been created, and they are now an essential part of the therapy of BRCA patients. Despite the fact that The Cancer Genome Atlas consortium has emphasised the significance of various omic data in breast cancer taxonomy, these tests concentrate on only a few genes' expression levels and none of them have been previously implicated in breast cancer.Some of them make use of the huge "omic" data that high-throughput technologies have enabled. High-throughput technologies like RNA sequencing (RNA-seq) have made it simple to collect a wide range of omic data at ever-increasing rates, accuracy, and lower costs, including whole-genome gene expressions, microscopic analysis, and other molecular data for the same BRCA sample. This allows researchers to investigate variations and similarities across multiple omic layers and improve BRCA stratification.

1.1 DEEP LEARNING

Deep learning is a type of machine learning and artificial intelligence (AI) that replicates how humans learn certain sorts of information. Deep learning is an important part of data science, which also includes statistics and predictive modelling.. Consider a child whose first word is dog to better comprehend deep learning. The youngster learns what a dog is and is not by pointing to things and saying the word dog.

1.2 GENOMICS

It differs from 'classical genetics' in that it considers the whole complement of hereditary information in an organism rather than simply one gene or one gene product at a time. Next-generation sequencing technologies have dramatically increased genome sequencing speed, capacity, and cost. Furthermore, bioinformatics advancements have enabled the construction of hundreds of life-science databases and programmed that encourage scientific research. These databases gather and organize information so that it may be accessed, compared, and analyzed rapidly.

1.3 SEMI SUPERVISED LEARNING

It is an example of ineffective supervision. Acquiring labelled data for a learning problem typically requires the employment of a trained human agent (for example, to transcribe an audio segment) or a physical experiment. The cost of labelling may make large, totally labelled training sets prohibitive, while getting unlabeled data is quite inexpensiveOrganic labels are common in data, but the technique may entail a human expert who adds tags to raw data to indicate the intended qualities to a model (answers). A label, in simple terms, is a description of what a model is anticipated to predict.

1.4 VARIATIONALAUTOENCODER

The important thing to remember is that our encoder network produces a single value for each encoding dimension. The decoder network then attempts to replicate the original input using these values. A variational autoencoder (VAE) is a probabilistic representation of an observation in latent space.

2. LITERATURE SURVEY

Breast cancer gene expression patterns identify TUMOR subclasses with clinical consequences. ROBERT TIBSHIRANI et al. propose in this paper. The tumors were classified into three types based on gene expression differences: basal epithelial-like, ERBB2-overexpressing, and normal breast-like. Clustering using two distinct gene sets indicated that these groupings were quite robust: the first, a collection of 456 cDNA clones originally chosen to reflect intrinsic tumor traits, and the second, a gene set that was significantly related with patient prognosis. This group consists of 40 previously investigated and described tumors. In all, 85 tissue samples from 84 persons were analyzed. Tissue samples were stored at - 170? C or -80? C in liquid N2. Malignant cells were found in more than half of the tumor tissues analyzed. In a prospective experiment, 51 patients with locally advanced breast cancer were treated with doxorubicin monotherapy before surgery, followed by adjuvant tamoxifen in the case of positive ER and/or PgR status. The ER and PgR statuses were determined using ligand-binding assays, and the TP53 gene mutation investigation was carried out in the same manner as previously described. [1]an update on multigene prognosis testing for breast cancer-emergent clinical biomarkers FERNANDO SCHMITT et al. suggest this investigation. Multigene signatures give significant prognostic information, which is especially useful for cancer patients whose prognoses are equivocal based on clinical criteria and traditional immunohistochemical markers. Clinicians are now provided genetic tools to assist them plan adjuvant medications, such as determining whether to continue adjuvant endocrine therapy or suppress adjuvant chemotherapy in patients when the side effects are exceedingly detrimental or the treatment is virtually unneeded.. In this part, we study and report on the applicability of the molecular tests Oncotype DX, MammaPrint, Prosigna, Endo Predict, Breast Cancer Index, Mammostrat, and IHC4 in the suitable setting. Importantly, the prognosis for triple negative carcinomas is extremely variable, and emerging molecular markers that better comprehend this extremely heterogeneous subtype of breast cancer may improve outcomes. therapeutic management of the illness. Gene signatures derived from gene expression research, bioinformatic tools, and clinical trials can now assist clinicians in evaluating the absolute advantages of systemic adjuvant chemotherapy or adjuvant endocrine treatment extension. Furthermore, molecular pathology opens up new opportunities for personalized therapy for each patient. [2]

Breast cancer risk prediction based on intrinsic subtype JOEL S et al. The intrinsic categories were significant for prognosis as independent. The combined model's C-index estimate (subtype and tumor size) beat either the clinicopathologic model or the subtype model alone. With a negative predictive value for pCR of 97%, the intrinsic subtype model predicted neoadjuvant chemotherapy efficacy. Diagnosis by intrinsic subtype complements recognized prognostic and predictive markers in breast cancer patients. The risk score and subtypes can also be used to predict neoadjuvant treatment effectiveness.. [3]

Prosing breast cancer gene signature test based on pam50 development and validation According to BRETT WALLDEN et al., Prostigmata, a PAM50-based subtype classifier and risk model on the Nano String nCounter Dx Analysis System designed for decentralized testing in clinical laboratories, is described. For each of the intrinsic subtypes of breast cancer, 514 formalin-fixed, paraffin-embedded (FFPE) patient samples were used to train prototype centroids on the Nano String platform. Using hierarchical cluster analysis of gene expression data, the prototype centroids established in previous PAM50 algorithm training operations were found. In the absence of adjuvant systemic therapy, 304 FFPE patient samples from a well-annotated clinical cohort were used to build a subtype-based risk model (i.e. Prosing ROR score). Prior to initiating clinical validation studies, the algorithm's prediction accuracy was validated using 232 samples from a tamoxifen-treated patient cohort.. The Prosing Risk of Recurrence (ROR) score model was demonstrated to be substantially associated with prognosis as a continuous variable, as well as to give much more information than both commonly used IHC markers and Adjuvant! Online. The training and validation data sets show that the FDA-approved and CE-marked Prosing test accurately estimates the risk of distant recurrence in hormone receptor positive breast cancer and can also identify a tumor's intrinsic subtype, which is consistent with [4]

Detailed molecular profiles of human breast tumors According to XYANG et al., provided significant insights into previously recognized gene expression subtypes and suggested the presence of four major breast cancer classes when we merged data from five platforms. Somatic mutations in just three genes (TP53, PIK3CA, and GATA3) were found in 10% of all breast tumors; nevertheless, there were other subtype-associated and novel gene modifications, including the enrichment of specific GATA3, PIK3CA, and MAP3K1 mutations with the luminal A subtype. HER2/EGFR/phosphorylated EGFR signature within the HER2-enriched expression subtype. Many molecular commonalities were discovered between basal-like breast tumors and high-grade serous ovarian tumors, implying a shared genesis and therapy possibilities. [5]

A REVIEW OF MICRONNAS and its Emerging Role as a Breast Cancer Diagnostic Marker According to Show kat Ahmad Bhat et al., breast cancer causes significant mental burden on the affected woman during the investigation, diagnosis, and treatment process since the breast is considered as a symbol of beauty, sexuality, and motherhood. Current biomedical research studies have used serum proteins, natural oncogenic genes, and advances in molecular technology such as microarray and RNA/DNase to study numerous potential breast cancer (BC) biomarkers. Small noncoding microRNAs have recently emerged as important regulators in oncogenesis pathways and can be used as noninvasive clinical diagnostic markers. This unique noninvasive biomarker family is developing as a potential chemical for cancer detection and prognosis. Recent clinical research has discovered miRNAs in human fluids such as serum and plasma, which opens the way to using them as noninvasive biomarkers of disease, including breast cancerAs a consequence, we propose potential future early stage detection approaches for breast cancer, demonstrating the usefulness and therapeutic potential of miRNAs in breast cancer. [6]

Convolutional neural networks were used to classify histological images of breast cancer. As argued by Fabio A et al. in this study, the success of most traditional classification algorithms is contingent on adequate data representation. A significant portion of the work is devoted to feature engineering, a challenging and time-consuming procedure that employs prior expert domain knowledge of the data to build effective features. Convolutional Neural Networks (CNNs) are a type of deep, feedforward network that has received a great deal of attention from the research community and industry In this article, we conduct exploratory tests using the deep learning approach to classify breast cancer histological images from Break His, a publicly accessible dataset available at database. Finally, we investigate the application of fundamental fusion rules to join disparate CNNs, which results in a marginal improvement in recognition rates. [7]

Convolutional neural networks were used to classify histological images of breast cancer. In this article, we conduct exploratory experiments using the deep learning approach to classify breast cancer histological images from BreaKHis, a publicly available database. This approach aims to enable the use of high-resolution histopathology images from BreaKHis as input to existing CNN while avoiding model modifications that might result in a more intricate and computationally expensive architecture. Finally, we investigate the application of fundamental fusion rules to join disparate CNNs, which results in a marginal improvement in recognition rates. To exclude hand-crafted features, we provided a set of deep learning tests done on the BreaKHis dataset. We proved that we could adapt an existing CNN architecture, AlexNet, designed for categorizingcolor photographs of objects to identify BC histopathological images. We also proposed several training strategies for the CNN architecture, which are based on the extraction of patches obtained randomly or via a sliding window mechanism and allow us to deal with the high-resolution of these textured images without changing the CNN architecture designed for low-resolution images. In terms of accuracy, our experimental results on the BreaKHis dataset showed that CNN outperformed traditional machine learning models trained on the same dataset but utilizing state-of-the-art texture descriptors. It is also possible to examine strategies for picking representative patches in order to improve accuracy. [8] Using high-dimensional genetic data, a deep-learning technique to detecting cancer subtypes was developed. According to RUNPU CHEN et al., cancer subtype identification has the potential to dramatically improve cancer prognosis and provide personalized patient management. Deep Type is a deep-learning system that learns cancer-relevant data representation with cluster structure by combining supervised classification, unsupervised clustering, and dimensionality reduction. Deep Type was applied to the METABRIC breast cancer dataset, and its performance was compared to that of leading techniques. In this work, we developed a deep-learning-based algorithm for cancer subtype detection in order to overcome many technological limitations of earlier approaches. The innovative concept outperformed two commonly used approaches significantly in terms of both internal and external assessment criteria. As a result, the method is inapplicable to cancers for which only a few samples have been analyzed. However, the cost of sequencing is expected to decline considerably in the near future, allowing for the collection of more tumor samples. However, many more cancers do not yet have molecular subtypes. Other clinical parameters, such as tumor grade, may be utilized to guide the identification of cancer subtypes, and we have shown that our method works well in the presence of label noise. Such possibilities require more investigation. [9]

Deep learning for integrated analysis of multi-omics data from breast cancer subtypes. According to SOMNATH RAKSHIT et al., breast cancer is a lethal condition that occurs all over the world and has been shown to be the top cause of cancer in females. Its discovery remains a huge computational and biological challenge. However, because of their high dimensionality and complexity, NGS techniques produce a significant volume of data that is difficult to analyses. Thus, one of the most challenging challenges in medical science is the integrated analysis of multi-omics data. As a result, reducing dimensionality before assessing data is useful. This work shows that a deep learning strategy, such as stacked auto encoder, can help reduce the dimension of a high-dimensional multi-omics data collection. Furthermore, numerous genes implicated in breast cancer pathways are discovered, as evidenced by the deep selection of features and the Bonferroni adjustment of p-values from one-sample t-test findings. The findings indicate that transcription factors (TFs) such as BRCA1, TP53, and ATM are implicated in breast cancer pathways. As the study progresses, meta-heuristics may be used with the deep learning technique to find key genes. [10]

Breast cancer prognostic subtypes and survival prediction based on multimodal genomics data Cancer, according to MD. Furthermore, for each input type, we constructed unimodal and multimodal features and trained ML baseline models such as decision tree (DT), Naive Bayes (NB), K-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF), and gradient boosting trees (GBT). Finally, we present the final forecast based on the model averaging ensemble of the top three models. We proposed an MAE for predicting the survival rates of patients with various subtypes of breast cancer. The results of the experiments for subtype categorization are positive, especially when it comes to the ER and PGR status, which has a 0.93 F1-score derived with a combination of GE and miRNA expression data. Despite not being the greatest at predicting survival rates, the GE + miRNA expression combo data as input generated really good results in general. [11]

Genenames.org: 2019 hgnc and vgnc resources As proposed by BRYONY BRASCHI et al. The Vertebrate Gene Naming Committee (VGNC) was established in 2016 to provide uniform nomenclature to vertebrate species that do not have their own nomenclature committees. The VGNC was the first to provide names to over 15000 chimp protein-coding genes. We've now applied this strategy to other vertebrate species, discovering over 14000 protein-coding genes in cows and dogs, and over 13 000 in horses. Our gene family sites have been renamed 'gene groups,' and their total number has surpassed 1200, with more than half of all discovered genes now assigned to at least one gene group. This post provides a summary of our online data and services, focusing on the last two years of work. [12]

Dropout is a basic method of preventing neural networks from overfitting. Overfitting, on the other hand, is a serious problem with these networks. It is straightforward to imitate the impact of averaging all of these thinned networks' predictions at test time by using a single untinned network with lower weights. This significantly reduces overfitting and offers significant advantages over conventional regularization procedures. Dropout improves neural network performance on supervised learning tasks in vision, speech recognition, document classification, and computational biology, resulting in cutting-edge results on a wide range of benchmark data sets. On SVHN, ImageNet, CIFAR-100, and MNIST, dropout algorithms achieve cutting-edge performance. Dropout also improved the performance of standard neural networks on new data sets dramatically. [13]

Variational Bayes auto-encoding DIEDERIK P. KINGMA et al. may be made more efficient by fitting an approximation inference model (also known as a recognition model) to the intractable posterior using the proposed lower bound estimator. Theoretical gains are mirrored in experiment data. Stochastic Gradient VB (SGVB) is a novel variation lower limit estimator for efficient approximation inference using continuous latent variables. The theoretical advantages are matched in the experimental results. [14]

A comparison of variational autoencoders VAEs are deep latent space generative models presented by RUOQI WE et al. in this paper. VAEs have shown to be extremely effective in a variety of applications, including image generation, image captioning, protein design, mutation prediction, and language modelling, among others. The main idea underlying VAEs is to learn data distributions such that new relevant data may be generated from the encoded distribution. These proposed enhancements will strengthen the capacity to generate useful synthetic data. This data can be used for representation learning or to enhance classification in deep networks if there is currently insufficient training data if a certain class is underrepresented. We anticipate that this will assist to improve the state of the art and lead to research breakthroughs in adjacent disciplines. [15]

Deep conditional generative models are used to learn structured output representation. KIHYUK SOHN et al. introduced supervised deep learning, which has been successfully applied to a variety of recognition problems. Experiments indicate that our proposed approach beats deterministic deep neural network equivalents in generating a wide range of realistic structured output predictions via stochastic inference. Furthermore, the training methodologies recommended complement one another, resulting in high pixel-level object segmentation and semantic labelling performance on the Caltech-UCSD Birds 200 and Labeled Faces in the Wild dataset subsets. [16]

In addition to empirical risk reduction, HONGYI ZHANG et al. suggested large deep neural networks. In this study, are powerful yet have undesirable characteristics such as memory and vulnerability to adversarial situations. The mixup is a data-agnostic and fundamental data augmentation principle. Mixup is a vicinal risk reduction technique that trains on virtual examples generated by linear interpolating two random samples from the training set and their labels. Mixup needs simply a few lines of code and little to no computational effort to include into existing training workflows. We proved that mixup improves the generalization error of state-of-the-art models using

ImageNet, CIFAR, speech, and tabular datasets. In addition, mixup assists in the prevention of incorrect label memorizing, sensitivity to hostile situations, and instability in combative training. [17]
"How can I believe you?" explaining the predictions of any classifier MARCO TULIO RIBEIRO et al. state that despite widespread adoption, machine learning models are basically black boxes. Understanding the underlying reasons of predictions, on the other hand, is crucial in assessing model trust. When deciding whether or not to act on a forecast or install a new model, trust is vital. This understanding provides new insights into the model, which may be used to turn an untrustworthy model or forecast into a trustworthy one. Explanations' value is proved in novel experiments, both simulated and with human subjects. Our explanations assist users in determining whether or not to trust a prediction, selecting amongst models, improving an untrustworthy classifier, and determining why a classifier should not be trusted. [18]
Malacards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search In this study, NOA RAPPAPORT et al. present MalaCards, which contains a web card for each of 20 000 illness listings in six worldwide categories. As a result of and GO-terms. MalaCards' ability to interlink iMalacards is a combined human illness database with different clinical and genetic annotation and structured search capabilities. NOA RAPPAPORT et al. provide MalaCards in this study, which includes a web card for each of 20 000 ailment listings in six global categories. MalaCards' affiliation with the GeneCards Suite of databases augments this with additional gene-related sickness information like as pathways, mouse phenotypes, and GO-terms. The ability of MalaCards to interlink information from disparate sources, combined with its sophisticated search function, relational database infrastructure, and convenient data dumps, allows it to deal with its rich disease annotation landscape and facilitates system analyses and genome sequence interpretation.MalaCards employs a 'flat' disease-card approach, but each card is mapped to popular hierarchical ontologies and includes information about multi-level disease relationships, making it an ideal tool for disease representation and scrutiny. [19]
Mirtarbase 2018 update: a repository of empirically confirmed microrna-target interactions According to CHIH-HUNG CHOU et al., microRNAs (miRNAs) are short non-coding RNAs of 22 nucleotides that are involved in negative regulation of mRNA at the post-transcriptional stage. Previously, we developed miRTarBase, a database of experimentally confirmed miRNAtarget interactions (MTIs). This improved version includes target locations validated by reporter assay and available in the literature. Furthermore, different browsing approaches enhance user browsing of certain MTIs. Since 2011, several researchers have used Mirta Base to conduct useful research, such as improving target prediction accuracy, integrating miRNA-related datasets with web-based tools, and reconstructing miRNA-gene regulatory networks in diverse illnesses. [20]

3.COMPARISON ANALYSIS

Title	Techniques & Mechanisms	Parameter Analysis	Future Work
Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications	The intrinsic gene set of 456 cDNA clones, chosen to ideally identify the intrinsic features of breast cancers, was used to analyses the 78 carcinomas and seven nonmalignant breast samples using hierarchical clustering.	We used SAM (16), a recently established analytical approach, to look for genes that were associated with patient survival. For this analysis, 1,753 genes were employed. SAM computes a score for each gene that indicates	To increase the efficiency of breast carcinomas by applying various clustering algorithms based on changes in gene expression patterns acquired from cDNA.

		the strength of its link with survival.	
An Update on Breast Cancer Multigene Prognostic Tests—Emergent Clinical Biomarkers	DNA profiling and micro-RNA analysis will enable the construction of multivariate models, as well as the formulation and implementation of novel gene expression signatures.	While pathology and immunohistochemical markers continue to be used to guide therapy, the physician may be met with ambiguous results that necessitate additional testing.	the path for future therapy decisions, more precise patient staging, and disease genomic subtyping Furthermore, molecular pathology opens up new opportunities for personalized therapy for each patient.
Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes	Methods Microarray and quantitative reverse transcriptase polymerase chain were used to create a 50-gene subtype predictor. Data from 189 prototype samples' reactions Test sets from 761 patients (no systemic therapy) were evaluated for prognosis, and 133 patients were evaluated for prediction of pathologic complete response (pCR) to a taxane and anthracycline regimen.	The subtypes and risk score can also be used to predict the effectiveness of neoadjuvant treatment.	To improve on present breast cancer prognosis and treatment prediction in the future.
Development and verification of the PAM50-based Prosigna breast cancer gene signature assay	The prototype centroids created in prior PAM50 algorithm training activities were identified via hierarchical cluster analysis of gene expression data. 304 FFPE patient samples from a well-annotated clinical cohort were then utilized to train a subtype-based risk model in the absence of adjuvant systemic treatment.	The training and verification data sets show that the FDA-cleared and CE marked Prosigna test accurately estimates the risk of distant recurrence in hormone receptor positive breast cancer and is also capable of identifying a tumor's intrinsic subtype, which is consistent with the previously published PCR-based PAM50 assay.	To improve clinical accuracy and technological precision in detecting breast cancer in the future.

Comprehensive molecular portraits of human breast tumours	DNA and RNA were isolated. 800 patients were tested on at least one platform in total. Different numbers of patients were used for each platform, with the most patients accessible at the time of data freeze; 466 samples (463 patients) were shared by 5 of the 6 platforms (example).	The integrated molecular studies of breast carcinomas presented here greatly expand our knowledge base, resulting in a complete catalogue of putative genetic drivers of the most frequent breast cancer subtypes.	To increase the accuracy of breast cancer prognosis and treatment prediction in the future.
MicroRNAs and its emerging role as breast cancer diagnostic marker- A review	Current biomedical research studies have investigated various possible breast cancer (BC) biomarkers by employing serum proteins, natural oncogenic genes, and advancements in molecular technologies such as microarray and RNA/DNASeq.	Nonetheless, large-scale investigations are required, and several elements of experimental reliability must be evaluated before circulating miRNAs may be employed as biomarkers.	to establish when tumor-related circulating miRNAs may be detected in the circulation during illness progression and to improve future efficiency
Breast Cancer Histopathological Image Classification using Convolutional Neural Networks	Convolutional Neural Networks (CNNs) are a type of deep, feedforward network that has gotten a lot of attention from the research community and industry after achieving empirical success in tasks like speech recognition, signal processing, object recognition, natural language processing, and transfer learning.	Alex Net, which was created for categorizing color photos of objects, was used to categories BC histopathology images. We also proposed several strategies for training the CNN architecture that are based on the extraction of patches obtained randomly or through a sliding window mechanism and allow us to deal with the high-resolution of these textured images without changing the CNN architecture designed for low-resolution images..	Strategies for selecting representative patches to increase accuracy in the future.
Deep-learning approach to identifying cancer subtypes using high-	By utilizing the potential of deep learning, we offered an innovative way to	We applied the approach to breast and bladder tumors, which have well-established	Various deep learning algorithms are predicted to yield more accurate

dimensional genomic data	disentangling and removing irrelevant elements.	molecular subgroups and may thus be used to guide the identification of novel subtypes. However, molecular subgroups for many more malignancies have yet to be identified.	and robust cancer subtypes in the future.
Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin	We initially used six distinct "omic" platforms to describe each of the unique tumour kinds. The "Pan-Cancer-12" tumour collection is made up of 12 distinct cancers.	The categorization findings from each single-platform study revealed sets of 8 to 20 sample groupings, each with a good association with tissue of origin.	the accuracy of breast cancer prediction using molecular classification analysis
Deep Learning for Integrated Analysis of Breast Cancer Subtype Specific Multi-omics Data	NGS technology have accelerated genome sequencing. However, the data's vast volume and complexity make it challenging to manage. As a result, it is advantageous to decrease dimensionality before evaluating data.	Deep learning techniques such as stacked autoencoder have been shown to be beneficial in lowering the dimension of high dimensional multi-omics data. The performance attained with this design is superior to the other solutions demonstrated here.	As an extension of the research, meta-heuristics may be used with the deep learning approach to identify important genes.
Prognostically Relevant Subtypes and Survival Prediction for Breast Cancer Based on Multimodal Genomics Data	To accommodate each data source, we created a multiplatform network called Multimodal Autoencoders (MAE) classifier that utilized DNA methylation, gene expression (GE), and miRNA expression data.	Although it did not have the highest performance on survival rate prediction, the GE + miRNA expression combination data as input produced very good results in general. Several reasons are impeding overall research: I a little quantity of labelled genomics data, which is likely to violate individual patients' privacy.	In the future, we plan to build a more robust multimodal network, such as multimodal Convolutional-LSTM, to act as both a feature extractor and a classifier, and to train with an enriched number of samples from other sources to create an explainable deep architecture that will allow us to learn more about potential gene set biomarkers-based diagnosis.
Genenames.org: the HGNC and VGNC resources in 2019	The Vertebrate Gene Nomenclature Committee (VGNC) and the Human Gene	Our Multi-symbol checker tool now allows users to enter a list of gene symbols to see	We intend to develop our curator tools in order to support orthologous gene

	Nomenclature Committee (HGNC) are utilised.	whether they are now HGNC approved, and the results page now lets users to sort and filter data, as well as modify how it is displayed, and save the whole results table in a comma separated file.	curation in several species at the same time. To help in the curation of more intricate homology connections over a larger variety of vertebrates, we will be constructing a semi-automated procedure including the comparison of phylogenetic trees.
Dropout: A Simple Way to Prevent Neural Networks from Overfitting	Drop out technique and neural network.	Several regularization approaches for preventing overfitting in neural networks have been developed. Dropout is another method of regularizing neural networks. Using the MNIST data set, we compare dropout with several of these regularization approaches in this section.	Speeding up dropout is an interesting direction for future work.
Auto-Encoding Variational Bayes	We provide Auto-Encoding VB (AEVB), an efficient approach for efficient inference and learning that trains an approximation inference model using the SGVB estimator. The theoretical benefits are mirrored in the experimental outcomes.	Stochastic Gradient VB (SGVB) is a revolutionary variational lower limit estimator for efficient approximation inference with continuous latent variables. Using typical stochastic gradient methods, the suggested estimator can be easily distinguished and optimized.	Because the SGVB estimator and the AEVB method may be used to practically any inference and learning issue with continuous latent variables, there are several potential directions: I using deep neural networks to learn hierarchical generative structures
Variations in Variational Autoencoders - A Comparative Evaluation	VAEs are deep latent space generative models that have proved enormously effective in a variety of applications including picture production, image captioning, protein design, mutation	We have supplied a complete understanding as well as a comparative assessment summary of the differences in VAEs so that scholars may understand the core theory as well as the intuition behind the variations in VAEs.	We anticipate that comprehending the VAE model from the perspectives of variational optimization and information theory will be major research topics in the near future based on the analyses and comparison

	prediction, and language modelling.	Furthermore, on GitHub, we offer reference implementations for the various VAE versions.	assessment offered in this study.
Learning Structured Output Representation using Deep Conditional Generative Models	Based on the conditional deep generative model with Gaussian latent variables, we suggested stochastic neural networks for structured output prediction.	Modeling the multi-modal distribution of structured output variables is a critical research subject for improving performance on structured output prediction tasks.	In the future, performance in terms of segmentation accuracy, calculation of conditional log-likelihood, and display of produced samples will be enhanced and improved..
mixup: BEYOND EMPIRICAL RISK MINIMIZATION	To address these challenges, we offer mixup, a basic learning concept. Mixup, in essence, trains a neural network using convex combinations of pairs of instances and labels.	On ImageNet, CIFAR, voice, and tabular datasets, we demonstrated that mixup improves the generalization error of state-of-the-art models. Furthermore, mixup aids in the prevention of corrupt label memorizing, sensitivity to hostile instances, and instability in adversarial training.	Although the discussion of these paths remains hypothetical, we are thrilled about the possibilities that the mixup brings up and believe that our insights will be valuable for future development.
“Why Should I Trust You?” Explaining the Predictions of Any Classifier	We offer LIME, a novel explanation approach that learns an interpretable model locally around the prediction to explain any classifier's predictions in an interpretable and faithful manner.	The framework allows for the examination of other explanation families, such as decision trees; a comparison study with real users would be fascinating. One issue that we did not address in this study was how to do the picture choice step.	We would like to investigate theoretical characteristics (such as the optimal amount of samples) as well as practical optimizations (such as parallelization and GPU processing) in order to offer accurate, real-time explanations for any future human-in-the-loop machine learning system.
MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search	The section Aliases and Classifications reflects a strategy for illness name integration across often contradicting sources, resulting in successful annotation consolidation.	The comparison of MalaCards may be taken a step further: how frequently can the validity of a disease gene candidate be predicted based on its being in the same biological pathway as a	In the future, this technique might be expanded to topological computations such as the disease-gene network distance of two illnesses or two genes.

		recognised gene for the same illness? This type of 'guilt by association' logic is prevalent in NGS studies, particularly VarElect'sindirect mode.	
miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions	The partial matching algorithm used in the proposed SPBA-based technique successfully extracts the relationship between miRNA and its targets from publications. Several curators have authenticated all of the articles in this update. Through the SPBA-based screening process,	Over 8,500 experimentally supported publications on miRNA-target interactions have been curated by us. As a consequence, 10,000 MTIs supported by good evidence have been added. Our upgraded text mining pipeline and a set of patterns created automatically by PBA increase MTI extraction.	In the future, we will give a large collection of experimental data that will definitely provide a continuing and valuable addition to miRNA research linked to cancer processes, illness detection, and therapy.

CONCLUSION

In this paper, we studied a computational strategy for exploiting larger datasets provided by high-throughput sequencing technologies for the breast cancer subtyping challenge. To do this, we developed both supervised and semi-supervised machine learning models that could take use of the much expanded variable space to correctly identify unique breast cancer samples. In terms of other data types and their combinations (miRNA and CNA data), we observed that combining multi-omic data did not improve overall prediction accuracy, most likely because to the unusually large feature dimension in proportion to the number of samples available. However, on the combined multi-omic data, single-layer Feed Forward Neural Networks (FFNN) and Variational Autoencoders (VAE) proved to outperform LR.

REFRENCES

[1] T. Srlie et al., "Gene expression patterns of breast carcinomas define tumour subtypes with clinical implications," Proc. Nat. Acad. Sci. United States, vol. 98, no. 19, pp. 10869-10874, 2001.

[2] A. F. Vieira and F. Vieira Schmitt, "An update on multigene predictive testing for breast cancer-emergent clinical signs," Front. Med., vol. 5, no. 248, 2018.

[3] J. Parker et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," J. Clin. Oncology, vol. 27, pp. 1160-1167, 2009.

[4] B. Wallden et al., "Development and validation of the PAM50-based prosigna breast cancer gene signature test," BMC Med. Vol. 1 Genomics Art. no. 54, 8, no. 1, 2015.

[5] Cancer Genome Atlas Network et al., Nature, vol. 490, no. 7418 (2012), pp. 61-70.

[6] A. M. Sabhiya, B. Showkat, and H. "Micrornas and its Emerging Role as Breast Cancer Diagnostic Marker- A Review," Tehseen, Advances Biomarker Sci. Applied Technology, vol. 1, no. 1, 2019, pp. 1-8.

[7] F. L. S. Oliveira, C. Petitjean, and A. Spanhol, "Breast cancer histopathology picture classification using convolutional neural networks," by Heutte, published in Neural Networks, 2016. [8] R. S. Goodison, S. Chen, and Y. Sun, "Deep-learning approach for cancer subtype detection utilising

[9] S. I. Saha, S. S. Chakraborty, and D. Rakshit Plewczyski, "Deep learning for integrated analysis of multi-omics data from breast cancer subtypes," IEEE Region 10 Conference Proceedings, 2018, pp. 1917-1922.

[10] M. G. Wicaksono, R. Karim, I. G. Costa, S. Decker, and O. Beyan, "Multimodal genomics-based prognostic subtypes and breast cancer survival prediction," IEEE Access, vol. 7, pp. 133850-133864, 2019.

[11] Braschi, B. B., Gray, K. A., Seal, R. L., Tweedie, and E. Yates A. Bruford, Nucleic Acids Res., vol. 45, no. D1, pp. D619-D625, 2017. "Genenames.org: The HGNC and VGNC resources in 2017," Nucleic Acids Res., vol. 45, no. D1, pp. D619-D625, 2017.

[12] Hinton, N. G., Krizhevsky, Sutskever, and R. Srivastava Salakhutdinov, J. Mach., "Dropout: A Simple Way to Prevent Neural Network Overfitting," Jan. 2014, Learn. Research, vol. 15, pages 1929-1958.

[13] D. M. Kingma and P. Kingma, "Auto-encoding variational bayes," Welling, Proc. 2nd International Conference on Learning Representations, pp. 1-14, arXiv:1312.6114, 2014.

[14] C. Doersch's "Tutorial on Variational Autoencoders," arXiv:1606.05908.

[15] K. H. Sohn, X. Yan, and Sohn Lee, "Learning structured output representation using deep conditional generative models," Proceedings of the National Academy of Sciences, vol. 28th International Conference on Artificial Neural Processes and Systems, pp. 3483-3491, 2015.

[16] D. Zhang, Y. N. Dauphin, and H. M. Cisse Lopez-Paz, "Mixup: Beyond Empirical Risk Minimization," Proc. 6th International Conf. Learn. Representations, arXiv:1710.09412 (2018, pp. 1-13).

[17] M. S. Singh, Ribeiro, and C. "How can I believe you?" Guestrin inquires. ": Explaining any classifier's predictions," ACM SIGKDD 22nd International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, 2016.

[18] Nucleic Acids Res., vol. 45, no. D1, pp. D877-D887, 2016. N. Rappaport et al. "MalaCards: A combined human illness database with diverse clinical and genetic annotation and organised search."

[19] Nucleic Acids Res., vol. 46, no. D1, pp. D296-D302, 2018. C.-H. MiRTarBase update 2018: A database of experimentally confirmed microRNA-target interactions. Chou and colleagues

[20] A. Hoadley et al., "Multiplatform investigation of 12 cancer types shows molecular classification inside and across tissues of origin," Cell, vol. 158, no. 4, pp. 929-944,2014