Towards Precision Medicine for Cancer Patient Stratification by Classifying Cancer By Using Machine Learning

By ***Mithun Sarker***

Lamar University, Texas, United States

DOI: 10.55662/JST.2022.3301

ABSTRACT:

On average, a drug or a treatment is effective in only about half of patients who take it. These patients need to try several until they find one that is effective at the cost of side effects associated with every treatment. The ultimate goal of precision medicine is to provide a treatment best suited for every individual. Sequencing technologies have now made genome means ics data available in abundance to be used towards this goal. In this project, we will specifically focus on cancer. Most cancer patients get a particular treatment based on the cancer type and the stage, though different individuals will react differently to a treatment. It is now well established that genetic mutations cause cancer growth and spreading and importantly, these mutations are different in individual patients. The aim of this project is to use genomic data to allow for better stratification of cancer patients, to predict the treatment most likely to work. Specifically, the project will use a machine learning approach to classify cancer and suggest medicine. The whole work is divided into two parts, one is predicting cancer using several machine learning classification techniques and then suggesting medicine.

1. Chapter I

1.1 Introduction

Precision medicine has made significant progress in the treatment of cancer during the past few decades. Targeted therapies, which target and kill tumor cells only, are an important part of precision medicine. But the research and use of these medicines are fraught with several difficulties, one of which is the identification of tractable targets. Medication toxicity and drug

resistance are major issues. Another important obstacle to the development of novel compounds is the increasing clinical attrition rate due to biological activity and safety concerns (Delhalle et al., 2018). As a result, strategies that recover ineffective compounds or discover dangerous compounds before costly preclinical and clinical trials could have a significant impact. Artificial intelligence (AI) is a branch of computer science which has advanced capabilities to analyze and predict tasks that otherwise would need human brains. AI is a highly effective solution-oriented tool, such as predictive capabilities and data scalability. It also has the ability to integrate and dimension. Thinking about the fundamentals of their nature and/or transformation of large data into useful clinical knowledge based on the knowledge gained from model data sets. The ability to learn is enhanced by improving the accuracy of predictions by using specific measures of performance. Particularly, machine learning (ML) (also known as deep learning (DL)-based methods were recognized and were deemed to be essential for biomedical data analysis because of accessibility to health data and the rapid growth of analytics techniques.

AI is utilized today to aid in the automated extraction of data from medical records, and to summarize handwritten notes or electronic records from doctors, as well as connect health data and save the data in cloud-based scale (big database) ; Garvin et al. In 2018; Syrjala in 2018) AI is an effective instrument that is able to significantly aid in all phases of treatment for cancer which range from early detection, stratification and identifying the cancerous tumor's margins after surgery, and the effect to medications and therapies and tracking the development of tumors and the possibility for resistance to treatment acquired in the future, and in predicting the severity of cancer or metastasis pattern and the possibility of the recurrence of cancer (Pinker et al., 2018). Cancer is the leading cause of death throughout the world .Cancers can arise from a variety of organs like. Breast, lung, and kidneys, which exhibit distinct phenotypic characteristics like cells' surface markers, molecular modifications that exhibit different Apoptosis and growth rates based on the microenvironment of the cancer and the condition of the blood supply and the aggressive nature of the disease. Furthermore, cancer is characterized by multiple mechanisms of progression and the underlying cause of disease ranging from gene-epigenetics to an extensive range of metabolisms. The fact that cancer/tumor is highly heterogeneous regards to the degree of heterogeneity between tumors (cancers from various patients) and the heterogeneity within a tumor (within the same tumor)

creates challenges for treatment, detection and the likelihood of recurrence. Treatment decisions for cancer have to consider not only the various types of cancer as well as the progression of the disease , but also the health of the patient in particular and ability to respond to treatment. Certain kinds of cancers, like pancreatic and gastric cancers are only discovered once they have advanced in their stage and often have relapses that are repeated. Integration of "multi-omics" (genomics, epi-genomics, transcriptomics, proteomics, and metabolomics), and "non-omics" (medical/mass-spectrometry imaging, patient clinical history, treatments, and disease endemicity) data could help overcome the challenges in the accurate detection, characterization, and monitoring of cancers. AI can play an important role for the evaluation of vast and varied data sets, specifically using multi-omics or inter-omics strategies and data integration to provide a comprehensive molecular diagnostic process for disease that can detect new and dynamic markers that can aid in diagnosing and prognosis. They also provide precise cancer treatment that go beyond the standard method of determining disease severity, like differential expression and the classification of supervised. Advanced computational analysis can enhance the understanding and automated analysis of radiographs taken by patients with cancer which rely heavily on visual assessment and, therefore, are not the same in terms of evaluation. It is the Moonshot for Cancer Moonshot(sm) research Initiatives from the NIH National Cancer Institute aims to collect as much information about omics and non-omics as it is possible from different regions of the globe to create a global network to exchange and analyze information related to cancer (Bates, 2010). The project will assist in developing human tumor atlas, determine the response to conventional treatments, improve guidelines for the systematic prediction of cancer as well as treatments, and determine methods to overcome resistance to drugs to enhance:

- The current knowledge of cancer
- Enable new strategies/technologies for cancer characterization,
- The early identification of cancerous tumors, and
- Extend treatments to greater numbers of patients in a personalized method.

The massive multidimensional biological datasets (including the individual variation in functions, genes and environmental factors) created and/or compiled to support this project across borders require sophisticated computation, and AI definitely could be one of the major factors.

1.2 Machine learning in Radiology of Cancer

Early detection and classification of cancer is essential for ensuring a precise diagnosis and treatment with a curative purpose. The term "radiomics" is the place where data derived from algorithms that identify and analyze the elements of medical images permits the advancement in precision in diagnosis of cancer, as also prognostication and prediction of the clinical outcome. The advances made in DL made use of convolutional neural networks (CNN) is the method that is used to analyze and differentiate images from cancer ones and help in the accurate detection as well as classification for cancer.

DL together with CNN and its variants is utilized to detect and classify cancers in various organs. Numerous studies using DL with radiation-related images for input are detailed within the next. With a set of around 130000 dermatological images which have been trained, the CNN can offer a dermatologist-level classification of cancerous keratinocytes of melanoma. Based on an exhaustive study of 11 studies, CNN provides accurate detection of hepatocellular carcinoma through the recognition of certain characteristics that are present in computerized tomographic (CT) and magnetic resonance (MR) images. This AI system that utilizes DL CNN algorithms to detect soft lesions of the tissue and calcifications showed the highest degree of precision for detecting breast cancer, which is comparable to that of a radiologists for breasts on the average. With two distinct datasets to develop and test the validity of AI algorithm, we discovered that the accuracy of a DL-based algorithm used to detect breast cancer by mammography is superior to the accuracy of radiologists with experience and areas of curves of receiver operating characteristics (AUC-ROC) that is employed to determine the accuracy of the AI system exceeding that of the average radiologist by 11.5 percent. The results of these research studies demonstrate the effectiveness of AI employing DL CNN techniques to precisely detect cancerous organs in various organs (La Thangue & Kerr, 2011).

The research that will be conducted in the future for clinical trials is anticipated to enhance the accuracy and efficacy in the detection of cancer with AI systems. Multidisciplinary research efforts and coordination are needed to discover ways that DL-based models can be used in the clinical setting.

1.3 ML in Histological of Cancer

The examination of the microscopic histopathology of cancer using immunohistochemistry is the most commonly used method for diagnosing and classifying cancer. Since the advent of scanner technology that allows the digitization of whole slides of removed or biopsied tumor samples has enabled computer-aided analysis to improve the accuracy and effectiveness of the diagnosis. ML/DL for analyzing digital images of the histopathology tumors has proven as a feasible technique to improve the diagnosis of cancer, and for identifying tumor metastasis and lymph nodes and for predicting the impact of genetic mutations and their clinical outcomes.

Here are a few examples which show the significance of ML within digital images of cancer, in order to improve the accuracy of diagnosis. Using DL with CNN technique for analysis of biopsied tissue specimens, prostate cancer as well as micro- and macro-metastases of breast cancer in sentinel lymph nodes could be automatically identified without the need for immunohistochemistry. Similar to the use of DL-based techniques to build an CNN to distinguish tumors from normal tissues, metastatic breast cancer is detected automatically by looking at images of lymph nodes that have been biopsied. In addition, DL-based algorithms can be trained to analyze histopathological images and identify the possibility of a clinical or genetic mutation. The DL method could automatically recognize the lung cancerous tissues or squamous-cell cancer and the normal tissues of the lung. Furthermore, the trained CNN is able to accurately predict the most frequently affected genes involved in lung cancer (Nevins et al., 2003).

A novel method of using DL to create deep networks was employed to analyze digital microarrays of tissues of colorectal cancers from 405 patients and their pathological and clinical characteristics, as well as their outcomes. The findings of this study confirm that DL-based predictions of the 5-year life expectancy for a specific disease are superior to predictions made through a visual analysis of histology conducted by experienced pathologists. DL coupled with CNN techniques has demonstrated the ability to enhance digitized histopathology-based diagnosis. In order to apply it for the detection and classifying of tumors in medical practices, the efficiency of ML/DL for the examination of digital tumor histopathology must be evaluated and verified in the future clinical trials that will involve many patients. In addition, combining ML data from digital histopathology and other sources

of data like tumor omics could enhance the precision of predictive capabilities as well as accuracy.

1.4 Applications of Machine Learning and Deep Learning Techniques to Omics Analysis

Following the completion of the International Human Genome Project and the dawn of what's called the post-genome era, the need for the application of genomic information in medicine has grown. It is why “genomic medicine" which is also known as "genomic medicine" is currently being considered as a possible new type of medicine that can provide patients with the most effective treatments based on the genomic data . This Precision Medicine Initiative was announced by US President Barack Obama for cancer and rare illnesses. The initiative classified those suffering from cancer or potentially suffering into subgroups of patients suffering from morbidity and identified the most appropriate treatments and preventative strategies specific to each of the subgroups. This initiative had a massive influence on global medical policies and since its inception the field of precision medicine has become one of the main goals that research has pursued in the field of cancer. The present situation that is based on precision medical treatment is blighted by a myriad of issues that have to be addressed. One of the biggest problems is that despite it being possible to find the best anticancer medications that are built on TGP, the main focus is on the field of precision medicine but the proportion of patients who receive the most effective treatment isn't as high.

The rise in the number of patients that are benefited by the application in precision medical treatment is believed to be a problem. It is also believed that treatment decisions made based on the current TGPs alone isn't enough to address this issue. It is essential to incorporate the entire genomic sequencing (WGS) data, along with other omics information like epigenome, proteomic or metabolic data to allow multimodal analysis. Particularly, recent developments in epigenome analysis have shown that cancerous cells suffer from numerous epigenetic diseases as well as genetic diseases. Epigenetic issues have a significant impact on the nature of cancer at the beginning of development to the time of progress. In addition, because of the advancement and development of epigenetic treatments for medical disorders and epigenome

anomalies it is essential to comprehend the pathogenesis of cancer through an in-depth look at epigenetic information.

In particular, as we've already mentioned that deep learning technology share four distinct characteristics and they are expected to play an essential role when it comes to the study of big multimodal medical data:

- Multimodal learning: various kinds of medical information (e.g. epigenome, genomic, or the proteomic record) may be placed together and used as inputs
- Multitasking learning: a range of tasks can be learned at the same time by sharing parts of the model
- Semi-supervised and representational learning: create methods to construct representations of data from massive amounts of unlabeled data, which lets you learn from very small amounts of labeled data
- There is a method to identify higher order connections between inputs.

These characteristics have led to the growth of techniques for deep learning and machine learning in large-scale analysis of medical data. Particularly, methods for deep and learning are widely utilized to reduce dimensionality as well as feature extraction as a means of extracting vital details from large volumes of data and to classify patients based on the characteristics that are extracted. A way to reduce the amount of dimensionality using an auto encoder was recommended for large volumes of data from multi-omics of 105 and we've used this method to identify those with lung cancer. The study we conducted combined the RNA-seq data with miRNA expression data of The Cancer Genome Atlas (TCGA) and was focused upon lung Adenocarcinoma (LUAD) as well as clinical data. We carried out a multi-omics research study employing an auto-encoder. We were able to sub classify patients according to their survival (categorizing good and poor prognosis of lung cancer types). The classifier was developed making estimates of labels taken from subtypes of patient as well as the support vector machine (SVM) generated the highest classifier results, with the accuracy was 0.82 from the dataset of test results. Subtypes were utilized to classify genes based on their levels of expression in the RNA. The 25 most popular genes were studied to determine factors that influence the prognosis of patients (Torres & Grippo, 2018).

Furthermore, we also identified survival-related subtypes of lung cancer that are not small-cell based on six kinds of TCGA multi-omics data. The result was a category dubbed"the Integrated Survival Subtype, which contained six types of data that successfully differentiated the prognosis of good and bad kinds of patients suffering from lung cancer. This was the statistically significant difference. This also proved that it did not depend on histopathological classification. In addition, the predicted subtypes were able discern between high risk as well as low-risk patients. In previous studies, we have provided a unique multi-omics study that may accurately predict the outcome of those with lung cancer.

There are also a range of issues when applying deep learning for health genomics data like genomic data. The primary issue is that the size of the input data is at a minimum an order of magnitude greater than the amount of samples. The input is greater than the sample size . Another reason is that the contributions of each variable to the model are frequently difficult to comprehend because of the various nonlinear processes. Thirdly, genetic information might not be delineated in terms of an intrinsic structure. To tackle these concerns, we have developed an analysis method for genomics which employs diet networks that have been modified by introducing per-element input scales.

In order to realize precision medicine, understanding the impact of anticancer medications is crucial, and efforts have been made to determine the impact of large-scale omics information using deep learning and machine learning methods (Seyhan & Carini, 2019).

The BrutonIbrutinib, which is a Tyrosine Kinase, is a very efficient treatment for those with chronic lymphocytic cancer (CLL) But there is a lot of variation in the course of the disease. Rendeiro et al. sought to predict the response in response to ibrutinibtreatment by studying multidisciplinary omics data with clinical information applying machine learning techniques. They also found an expression pattern for one gene that corresponds to the general reduction in function of immune cells as well as the shift to an inactive state in response to treatment with ibrutinib. There was a distinct variation for each patient in the speed of execution for the treatment. This variation could be used to assess the specific patient's response to ibrutinib based on the pretreatment samples of the patients . This study revealed time-dependent cellular, molecular, and regulatory effects on the therapeutic inhibition of B-cell receptor

signaling in CLL using machine learning multi-omics analysis and established a widely applicable method for epigenomic/transcriptomic-based therapeutic monitoring.

Recently, the advancement of precision cancer therapies that is based on large-scale omics analysis using machine learning and deep learning techniques is currently being investigated but the field is only at its infancy. One of the problems being studied is the biochemical consequences of a mutation within a gene can differ based on the purpose of the gene and the location of the mutation. However we've been unable to take this particular information into the analysis. The different mutations are combined to be examined. In the near future , we need to think of ways of the integration of this knowledge into the analysis. Additionally, omics data, such as epigenomic and genomic data contain a variety of parameters in relation to the quantity of data samples. This makes it difficult to analyze the information in its entirety and therefore it is crucial. The most crucial factor is choosing the right model that allows an improvement in parameters without compromising the realism of the model (Rello et al., 2018). This is why more research is needed. In addition, recent studies have revealed key signaling and molecular pathways that are responsible for the growth of cancer and its progression. Various ways of analysis have been employed to understand the root cause of cancer and determine potential targets for treatment with the help of massive datasets. The methodology is proven and many research studies were published which can be significant for the advancement and development of research in oncology.

It is important to realize that limitations exist in the studies conducted with Dry lab methods and it is essential to validate the results obtained by this method through suitable wet lab research (cell-level research or animal studies using mice). When you transfer the findings of your wet lab research to the dry lab method, it is expected that the accuracy of results obtained using dry lab techniques will increase. This is true for AI everywhere in the world regardless of the advancements AI will achieve soon. It is extremely risky for humans to depend on AI's expertise in everything. We believe that there should always be a human-based verification process. In the ideal scenario, we should be aware of what are the strengths and weaknesses of AI and humans, which will allow them to collaborate and complement each other.

2 Chapter II

2.1 Literature Review

A brief overview of the available research on cancer-related studies that employ ML in the detection of cancer, prognosis, as well as classification of patients. In the sections that follow the well-known and promising techniques in DL and RL and their efficacious applications and effects on cancer research and in clinical practice have been reviewed. The clinical scenarios we have discovered through an overview of our study are discussed clearly , as well as their associated research in the sections below. These scenarios within the clinical setting are one of the most efficient areas of biomedical-based ML-based applications within the medical field. We examine the factors that determine robustness as well as the capability to describe the reasons for using ML-based models to aid in healthcare. A summary and outlook of the most recent developments and future challenges in the use of models based on ML to automate the process of making decisions within the practice of medicine is included (Parimbelli et al., 2018).

2.1.1 Deep Learning in Research on Cancer

The value of AI/ML strategies in the area of biomedicine and precision oncology are being realized because of the advances in the use of advanced algorithms for diagnosing cancer using computers. These cutting-edge technologies are being integrated into the routine of the clinical practice, in the hope of improving the quality of care provided to patients while speeding the decision-making process in clinical settings. DL methods, a kind of ML, is providing enormous assistance to doctors in the area of medical oncology due to the creation of diagnostic medical imaging systems that seek to improve the diagnosis of illnesses and the early diagnosis of cancers. Because of the enormous amounts of data and the parallel and

distributed frameworks for analyzing the data, DL structures have been developed and can be classified into three categories:

- Deep neural networks (DNNs)
- The Convolution Neural Networks (CNNs)
- Recurrent neural networks (RNNs)

DL can be described as artificial neural networks which have numerous non-linear layers. The primary distinction with DL is the fact that their feature layer was developed by using the general-purpose method of learning instead of being constructed through input from the user.

ML is broadly divided into three distinct models:

- The task with a label class,
- The unsupervised task in which no label or class is provided
- The most advanced kind of RL methods in which the agent is trained to carry out actions in a series.

These techniques are applied primarily in the application of ML to automate decision-making for patients, using Decision Trees (DTs) as and the Support Vector Machine (SVMs) and linear regression being among the most frequently used techniques. Based on conventional ML techniques, the main variables or descriptors are utilized to build a model and then discover patterns, and suitable definitions of feature vectors relevant to the subject being studied. However, the capacity of traditional machine-learning techniques to examine the raw data from natural systems isn't nearly as efficient. But, today, DL (i.e. the use of neural networks using multiple layers) is receiving a lot of interest due to their capability to include multiple layers of representation of the characteristics involved in the process of learning, which improves the effectiveness of the model, as well as its ability to scale computationally and its feasibility (Chiu et al., 2020). The DL approach is able to be tailored to various representations of data that allows different types of traits to be drawn from the more relevant data employing a sophisticated learning process. DL is superior to other methods that involve perception, and the techniques for ML have difficulty managing massive amounts of data.

In the area of cancer research in oncology and medical research, various DL designs have been developed and used for the evaluation and/or identification of various kinds of cancers .The DL models have been utilized to identify and classify cancerous forms. The study of DL models' performance has shown that these techniques are superior to other methods for ML. Frameworks for DL have also been designed and applied to aid in the detection of cancer and classification through analyzing the gene expression. Regarding the prognosis for cancer, as well as treatment choices, DL methods have been designed to resolve the problem of predicting responses to treatment in certain cancer types (Osman, 2019).

2.1.2 Learning Reinforcement in Research on Cancer

RL represents a distinctive type of ML. It has discovered applications in cancer research and medical oncology for determining the most effective treatments options and computer-aided diagnosis. Within the RL model, an agent (i.e. the doctor) learns by interacting with their environment in order to reach goals that are based on a result that the agent hopes to attain (reward the role). The process of learning for an agent in the normal RL process is a continuous process. Interactions with the surrounding environment happen at various time intervals. Once the state of an environment is recognized the agent decides to take the action that interacts with it. The environment responds to the action chosen and the reward that the agent receives or gets is determined.

Integration of DL as well as RL systems into the clinical setting and with a focus to the existing medical and non-biomedical cancer information will improve comprehension of the complexity and depth of cancer as well as the importance of the identification of risk factors as well for determining the basis of effective treatment strategies.

2.1.3 Diagnostic and Detection of Cancer

It is true that automated cancer diagnosis and detection is among the most vital and profitable fields of medical applications using ML. In our findings ML-based methods based on traditional or advanced techniques to carry the diagnostics for common cancers like colon, breast and pancreatic cancers and many others. Most studies used images generated by computed-tomography (CT) in addition to magnetic resonance imaging (MRI) as well as radiography using X-rays, as well as the positron emission tomography (PET) to develop automated diagnostic models built on DL structures. In this study the most recent and

relevant studies of this year used either imaging, genomic or other relevant clinical information to develop models employing ML to assist in the diagnosis and detection of disease. A significant part of this list of studies is due to studies which focus on particular clinical conditions using images as inputs for models to study DL (Prados et al., 2015).

In this context the automated detection of disease was studied making use of CNN models that could recognize early cancers of the breast with Histopathological scans.In particularly, Zheng et al.evaluated and suggested using a transfer technique based on CNN that detects early signs of breast cancer by effectively splitting the ROIs. When compared to other methods of machine learning and methods, the results were impressive that were achieved with high levels accuracy (i.e. 97.2 percentage) and a balanced ratio between sensitivity and specificity metrics (i.e. 98.3 percent as well as 96.5 percentage (or 96.5 percent and 98.3 percent and 98.3%, respectively).

Based on the algorithm employed for preprocessing, training and evaluation, satisfactory results were found with an average accuracy of was close to 90.0 per cent, and is a testament to the contributions of the authors to assist clinicians in their diagnosis procedures. Frameworks that support DL were also designed and developed within and after the CNN structure to analyze CT and dermoscopy images of liver and skin cancers as well as skin cancer, respectively. Conventional ML algorithms such as DTs as well as random forests (RFs), Naive Bayes (NB) and K Nearest Neighbor (kNN), Artificial Neural Networks (ANNs), Gradient Boosting Machines (GBMs) and SVMs were also utilized this year's medical oncology for the automated detection and diagnosis of cancer. The studies that can be considered indicative include those that showed positive results. The results were achieved by using traditional ML methods to analyze the clinical as well as laboratory, genomic and epidemiological information to detect cancers of the prostate and lung, stomach, and breast cancers. In a different study, methods that were unsupervised and supervised utilized to analyze transcriptomic data to determine biomarkers that may be biomarkers that could be candidates (i.e. genes) which may be associated with the development of pancreatic cancer. Bioinformatics-based preprocessing workflows were applied to identify the latest gene set involved in the spreading of tumors of the prostate with an AUC greater than 0.90. Curve (AUC) higher than 0.90.

2.1.4 Cancer Patient Classification

In the area of medical oncology and studies on cancer, the classification method of diagnosing disease has been meticulously investigated using the long-running ML techniques to address the issues of binary or multi-class learning. The categorizing of cancer patients in groups may help in the creation of models that predict disease with ML that can be used to analyze risk stratification with the capability to broaden the performance. In this regard, a number of research papers were released last year, which highlighted the determination of the most crucial factors in determining the severity of cancer applying conventional methods and DL[9]. The majority of these research studies employed DL models to study the genomic and imaging data regarding the stratification of risks and their prediction. The models that were trained by DL were able to detect and classify diseases using photographs and genetic data. Data-driven approaches demonstrated the benefits of ML-based frameworks to utilize heterogeneous sets of data for better diagnosis and treatment (Vargas & Harris, 2016).

Recently, a fascinating study was proposed by Li and colleagues with respect to the study of tumor-specific Ras pathway activation and the identification of hidden key players as the cancer develops. The DNA sequencing, copy number and mutational information were incorporated in the DL model to gain insight into the functions that the pathways perform. The model was superior in performance when compared to other studies that focus on the detection of abnormalities in the Ras pathway in cancer-related samples (i.e. an AUC greater than 0.90) In a different study (CRC) group was examined by using whole-genome-sequencing experiments of DNA samples in order to develop an ML model that has accurate generalization capability for the early detection of cancer. A complete ML-based strategy was devised to examine the genetic profile and stage of cancer, and to find the covariates with high scores that differentiate between early-stage CRC and control patients. There are also well-known adaptive ML algorithms that have been extensively utilized in the literature to aid in the detection of cancer by integrating different kinds of data.

Song et al. created a predictive model utility to predict the future results of bladder cancer. This model relies on the capacity to understand ML algorithms. The classification model was developed by combining clinical and molecular aspects that are capable of distinguishing cancerous and non-cancerous ones. Positive results were obtained on the effectiveness of classification with an AUC of more than 0.70. Recently, similar research was published with

the intention of using methods that are based on data to categorize cancer-related data for the prediction. These studies could be compared with ML-based models that assist in making medical professionals make decisions when it comes to monitoring and follow-up for patients. Because of the availability of many kinds of data within the realm of research into cancer the studies utilized cancer information from the patient registers as well as electronic medical records, as in addition to sequences, demographics and imaging technology (Regel et al., 2020).

Two distinct studies have been published which employ CT data which is then paired with radiomics to classify cancer patients to aid in the detection of cancers in the lung and lung lesions. Integration of radiomic and clinical data in the context of ML algorithms enabled the identification of specific characteristics of patients that should be considered when screening for illnesses. The performance indicators that were associated with the proposed ML-based strategies were very good, with high accuracy in classification and an AUC greater than 77.0 percent , and 0.80 and 0.80 respectively.

2.1.5 Prognosis for Cancer and Prediction of Survival

It's an entirely different area of research into cancer that AI will provide important insights regarding the treatment of patients diagnosed with cancer. In particular, in this area we have collected research which aims to establish the prognosis of patients i.e. determine the probability of survival based on a mix of variables (clinical imaging, genomics and clinical) and analyze the responses to treatment, and ultimately, the prognosis of patients. Due to the volume of information available and the complexity of its nature, these studies aren't possible without the help of ML algorithms, specifically DL techniques. In the last year, over 200 scientific studies have been published in order to establish the prognosis for cancer. In all the studies, the majority of them used DL techniques, but only a handful used traditional methods for ML[10].

Similar to the earlier situations, in line with the general research on cancer, specific organ cancers are most studied, which include prostate or lung cancer, breast cancer, and colorectal. The types of data used vary from research to research, however there is a trend towards specific sources of information for every type of cancer. In particular, it is the pathology data used in prostate cancer research as well as breast cancer research, and also colorectal cancer.

Research that uses genomic data as well as lung cancer research is heavily dependent on the imaging information, especially CT scans. Despite the low propensity for each type of cancer it is evident that all ML methods including DL techniques are mainly employed to analyze images regardless of the type of imaging technique used. A fresh approach was revealed, where an automated deep-learning system was programmed to recognize prostate biopsies with the Gleason system to grade. Similar techniques have been reported in the literature to evaluate the prognosis for prostate cancer . The same manner, a method based on DL has been proposed to differentiate between benign and malignant skin lesions, which gives an overall AUC that is 0.91. Another use for ML when it comes to cancer prognosis is to estimate the length of time a patient will live by using several factors that are determined from the base. By incorporating data from subsequent follow-up visits improves the accuracy. The findings of these studies can be found in literature on various cancer types, e.g. lung cancer, breast cancer, bladder cancer, etc. Like the purpose, ML algorithms also have been used to predict the reaction to treatment, and, as a result in assessing the prognosis of the patient as well as chances of the patient's survival (Piñeiro-Pérez et al., 2022).

2.1.6 Robustness and the ability to explain AI/ML models

Recent advances in AI/ML raised to light the issue of vulnerabilities that impact predictive models and impact their reliability. In this context, guidelines that guarantee the security and credibility of ML-based methods in today's digital world. These guidelines are being developed to encourage technological advancement while safeguarding Human Rights. While ML-based methods can discover complex patterns and connections from massive data sets however, there's a deficiency of information about causal connections and the precise guidelines.

To ensure proper usage of models using ML in clinical practice , according to a suitable regulatory framework, three main concerns must be analyzed and addressed. The first one is the model's transparency , which is linked to the technical requirements as well as the data utilized to be achieved. To have a complete view of the ML model, it is necessary to understand the different levels that it can be applied (i.e. fundamentals of technology) and the specifications (i.e. specifics regarding the testing and learning phases) as well as the capability to understand (understand the logic of the model) is required. The third aspect concerns the validity of models and also the technical options that need to be formulated and implemented

in order to prevent the failure of autonomous systems in certain conditions. To evaluate the legitimacy of a model, its capabilities and weaknesses need to be evaluated. A lack of results or problems indicates that the model used for ML isn't reliable. Thus, the strategies comprise:

- Cleaning the data
- Solid learning and
- Rigorous testing has been suggested to increase the accuracy of ML models.

Security of data when using an ML system is the third aspect that needs to be taken into consideration in order to establish the appropriate legal framework for creating and automating systems. Implementing guidelines on data protection ensures the system is in conformity with privacy and security laws. However, using methods of anonymization and pseudonymization of sensitive data, according to the General Data Protection Regulation (GDPR) in Europe and the guidance on how data should be used or shared in accordance with the regulations of the Health Insurance Portability (West et al., 2006).

Understanding the mechanisms and reasoning of the mechanisms and reasoning behind an ML system in the digital society will guarantee its security. Introducing standardized approaches to assess the robustness of predictive models with respect to the data used for training, promoting model's transparency through explainability-by-design principle for ML-based systems and designing methodologies to address vulnerabilities ensuring thereby the reliability will promote an effective and secure use of AI/ML systems. Furthermore, the implementation of best practices to ensure the correct development and implementation of automated ML-based platforms will provide the legal framework to improve the faith in AI systems.

The idea of Explainable AI (XAI) provides a conceptual framework to aid in understanding that the AI machine or system can achieve a specific outcome. The process of analyzing the outcomes of a model as well as providing explanations on both the local and global levels can enhance the effectiveness of systems that provide clinical decision-making aid. Examining models which are model-neutral that are specific to the specific model may be used to explain black-box models' output , increasing their reliability and transparency in clinical research. Model-specific explanations are very popular but aren't easily adapted to two different models. If a novel design for a predictive model is discovered an innovative method of

analysis and diagnosis of the model needs to be considered. But models that are not non-model-specific can improve the capabilities of exploratory models by employing instance-level exploratory techniques that provide a deeper understanding of how the model is able to make an estimate for just one observation. Apart from instance-level explanations, dataset-level-explainers for ML-based predictive models help to understand how the model's predictions perform for the entire dataset and not for a certain observation. Regarding the explanations given by models that are based on networks, and trees-based classifiers.While DL techniques have been demonstrated to be extremely efficient and effective in their efficacy and efficiency, explanations about why the DL model arrived at outcomes should consider more detailed methods based on models specific and models-agnostic analysis[11].

3 Chapter III

3.1 Methodology

3.1.1 Method

Our models were trained using Merck and Co.'s oncology screening (O'Neil et al., 2016). The collection includes 10340 cancer samples and 713 normal tissue samples. Twelve experimental and twenty-one licensed anticancer medications (see Supplementary Table S2) were used to build pairwise combinations, whereas the remaining eleven (the "supplemental" set) were only evaluated in conjunction with the exhaustive set of 20. (see Fig. 1d). Using a 4-by-4 dosing regimen for each sample, cell growth rates were assessed after 48 hours and compared to the control. The checkerboard was omitted in favor of single detached monitors with eight concentration levels and six replicates, which allowed interpolating values collected from fitted Hill curves to define the edges of a composite surface (where a drug is absent). This produced 5-by-5 concentration point area for each sample. (see Fig. 1c).

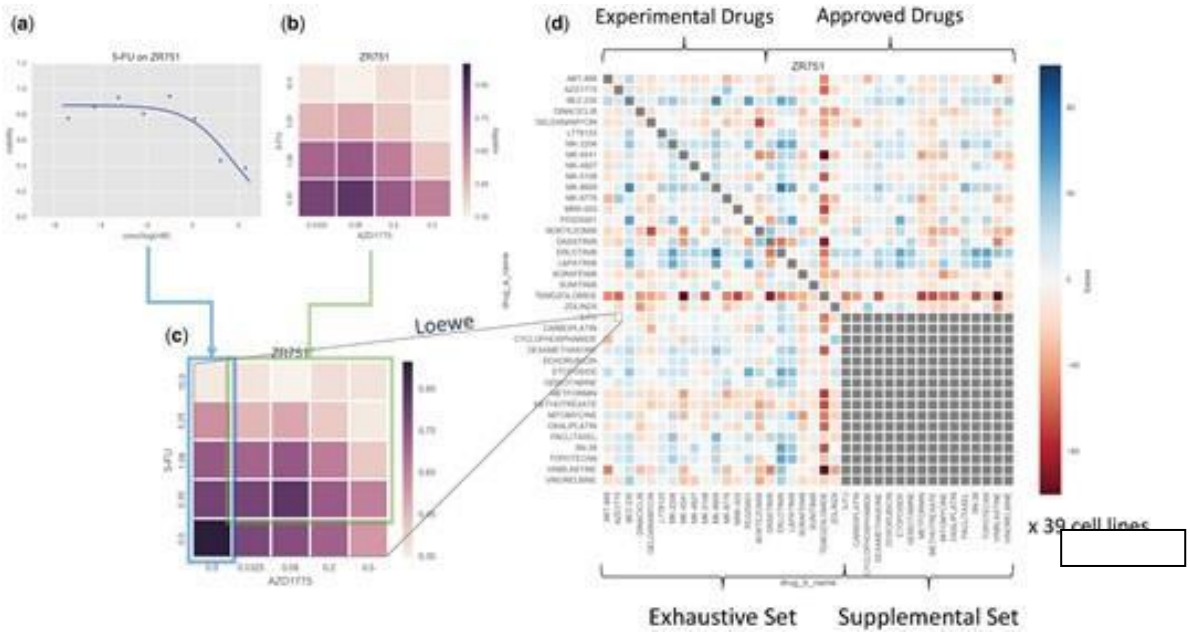


Figure 1

3.1.2 **Models**

- CNN
- SVM
- Naive Bayes
- KNN
- Random Forest

3.1.3 Model Details

One-dimensional kernels are applied to the vector of gene expression in this CNN model. An FC and prediction layers receive the 1-D convolutional layer's output before being passed on to the maxpooling and prediction layers (Fig. 1a). This model is referred to as a 1D-CNN for simplicity's sake. For time series prediction, the fundamental difference between the proposed 1D-CNN and conventional CNNs is that the stride of the convolution is equal to kernel size. 1D CNN is used in some cases to capture the temporal relationships between nearby values in the input data set. As a result, we set the CNN's stride as large as the kernel size in order to capture only the kernel's global properties, rather than the correlations between surrounding gene expression data.

3.1.4 Data Preparation

- "After the dataset is loaded from the pckl files, it's then separated into 2 arrays" "(X_s for the Data itself without the labels, y_s for the labels only)"
- "Reshape the X_s to be in 2d format instead of the 4d format to be suitable for the ML models (of shape (n_samples, n_features)"
- "Encode the y_s to be numeric values"
- "Split the dataset into training and testing (split_size = 0.3)"

3.1.5 Implementation Plan

- "Analyze the dataset".
- "Taking care of missing data"
- "Clean up data"
- "Classify data using several classification algorithms"
- "Find correlation between genome pattern and treatment being used Table of Content"

Step

After collecting frozen dataset (pckl) file , we used CNN to predict cancer type. I used the following network architecture

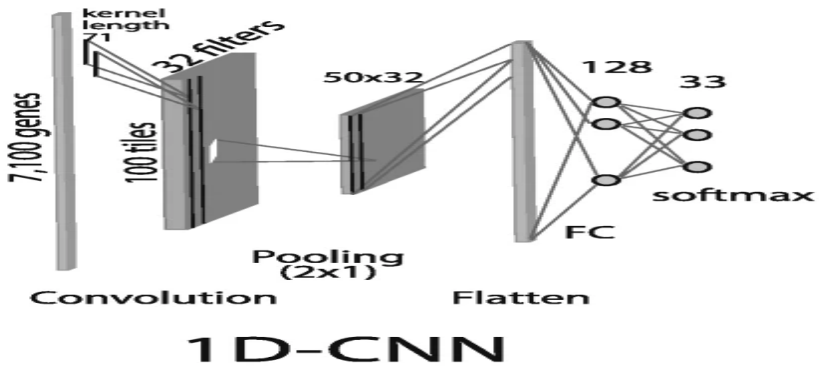


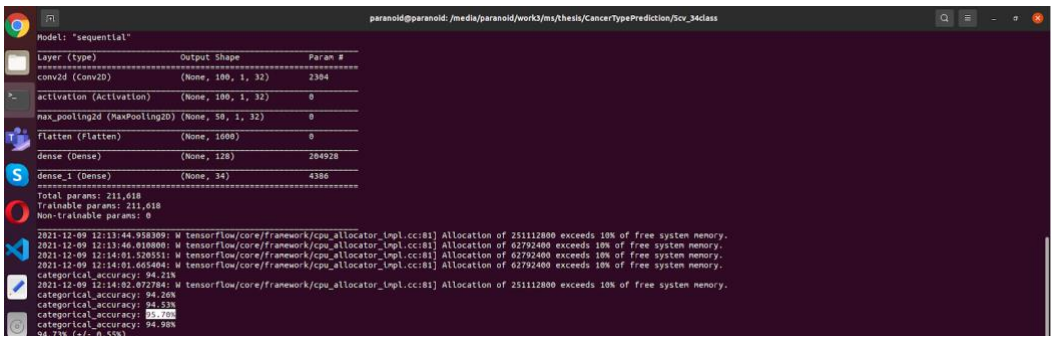
Figure 2

3.1.6 Hyper Parameters

layers	**[8182,4096,1]**	**Dropout**	**0.5**
epochs	**1000**	**Input-dropout**	**0.2**
Act_func	**Tf.nn.relu**	**eta**	**0.00001**

3.1.6.1 *Parameter Tuning*

Epoch 100 and Batch Size :256 , 10 folds : Mean accuracy :94.13 Epoch 100 and Batch Size :256 , 10 folds activation function sigmoid on last layer : Mean accuracy :93.29



3.1.7 Comparison of Accuracy of Several Models

Model	**CNN**	**SVM**	**Naive Bayes**	**KNN**	**Random forest**
10 fold	Min:93.58 Max:95.66 Mean:94.92	Min:94.56 Max:96.65 Mean:95.68	Min:86.42 Max:88.43 Mean:87.53	Min:93.48 Max:91.13 Mean:92.18	Min:92.71 Max:94.02 Mean:93.17
5 fold	Min:94.42 Max:95.75 Mean:95.13	Min:95.11 Max:96.52 Mean:95.49	Min:87.38 Max:87.78 Mean:87.52	Min:91.54 Max:92.30 Mean:91.98	Min:92.71 Max:94.02 Mean:93.17

3.1.8 Result Comparison in RED PC and Local Machine

Algorithm Name	**Red PC**	**Local Machine**
CNN	**38.18**	**33.79**
SVM	**847.48**	**411.14**
NAIEV BAYES	**847.48**	**411.14**
KNN	**1831.33**	**28.33**
RNN	**166.38**	**165.19**

3.1.9 Improving Cancer Medication Efficiency Using Synergy

3.1.9.1 Synergy score

Using theoretical models like Loewe Additivity, Bliss Independent (Bliss, 1939), Single Highest Agent (Tan et al., 2012), or a more recent Zero Interactions Potency, researchers may estimate how much synergy a surface has (Yadav et al., 2015). Only raw surfaces were released as extra material, as mentioned in the original paper.Consequently, we employed Combenefit's batch processing mode to create Loewe Additivity values" (Di Veroli et al., 2016). A total of 10340 quartets were created as a result of this process (compound, compound, cell line, synergy value).

3.1.9.2 Molecular Characteristics and Chemical Descriptors of Cancer Cell

A combination of pharmacological and disease genetics was employed to generate numerical representations of input data. Open Babel was used to protonating the chemical representations to a pH of 7.4 after salts were removed. The chemistry of the two drugs in the drug combination was then analyzed and analyzed in detail. Calculations were made for three distinct categories of chemical properties. Extended connection fingerprinting with a 6-degree radius was generated using jCompoundMapper (ECFP 6). Cao et al., 2013) was also used to calculate pre-set physicochemical characteristics. The collection of chemical characteristics was finished with the acquisition of binary toxicophore features based on a set of toxicophores from the literature. Toxic structures are known as toxicophore. As zero variance traits were eliminated, a smaller range of chemical features could be considered. The final collection of chemical characteristics includes 1309 ECFP 6, 802 physicochemical, and 2276 toxicophore properties. .Gene expression patterns were used to characterize the cell lines. An untreated cell profile was supplied from the Array Express database (accession number: E-MTAB-3610). Affymetrix performed the measurements on the Human Genome U219 arrays plate. Analysis of Microarray Factors Summarization was used to summarize the quantile-standardized original data (FARMS). Talloen et al. (2007) employed FARMS' Helpful calls for every gene to filter down the gene expression data to something like a final collection of 3984 genomic characteristics for analysis.

3.1.9.3 Deep Synergy

It is a feed-forward neural net that transforms input variables into a single output value. Concatenated vectors combine the features of two medications and one cell line to characterize

the samples. Deep Synergy is shown in its most basic setup in Figure 2. The neurons inside the input layer receive gene expression profiles and chemical description for both medications. The output unit analyses the data as it travels through Deep Synergy system and generates the projected synergy score. We present each sample twice in the testing phase since the network should not be able to detect the distinction between the pharmacological combination AB supplied in the sequencing A-B or B-A. Both A-B and B-A sequences of the drug properties are used. Both sample representations are disseminated around the system and averaged for predictions. No matter how the pharmacological combination AB is taken, Deep Synergy can accurately estimate its value.

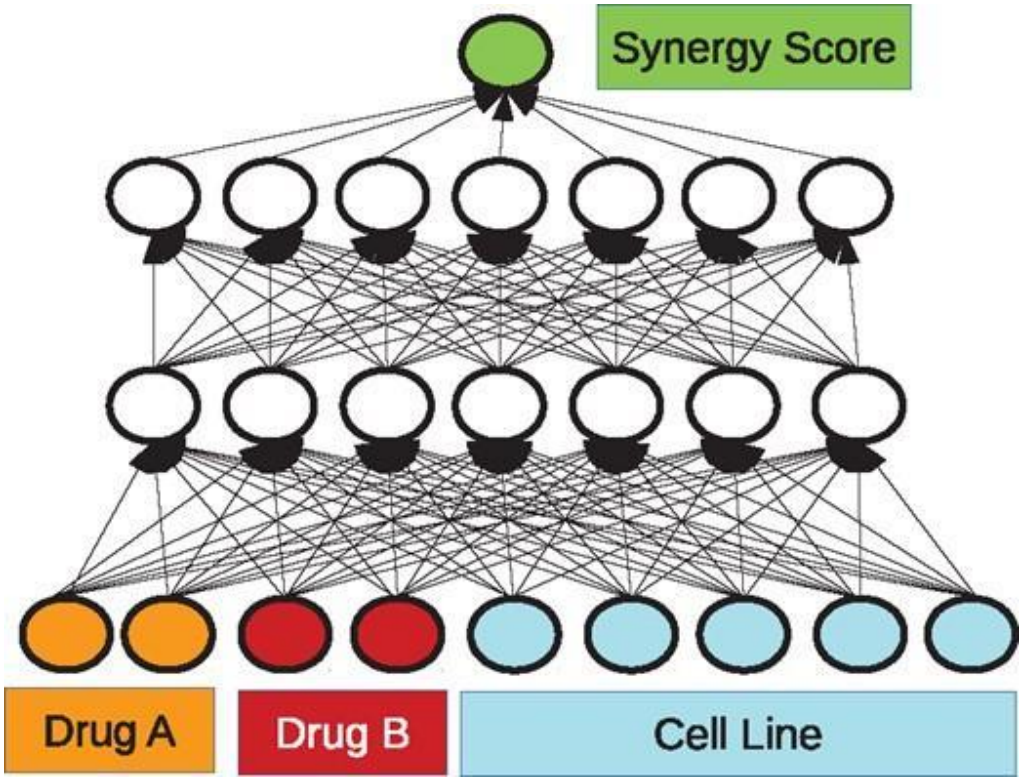


Figure 3

A variety of different data normalization methods were tried while working with different numbers of neurons and hyper parameter values in conic or rectangular layers. Other techniques of regularization and learning rates were also studied. The hyper parameter space is further explored in the sections that follow. Table 1 provides an overview of the hyper -

parameter space under study. The hyperparameter selection technique, whose findings are shown in Supplementary Table S7, was used to design the architecture of DeepSynergy. Tanh normalization, which consists of first standardization, then a hyperbolic tangent, and finally a second standardization, was found to be the most effective. Conic layers are also included in DeepSynergy. The fact that conic layers function well might be explained by their regularizing impact. The model is forced to generalize by creating just the most important representations of chemical characteristics of the input compound combination due to the decreased number of parameters accessible in the upper layers. Furthermore, a large number of units (8192) in the first layer performed better. For learning performant networks, a lower learning rate (105) and dropout regularization were also required. DeepSynergy has a conic architecture overall.

Overall, DeepSynergy has a conic architecture with two hidden layers having 8192 neurons in the first and 4096 in the second hidden layer. It uses tanh input normalization, has a learning rate of 10−5, an input dropout rate of 0.2 and a hidden layer dropout rate of 0.5.

	drug_a_name	drug_b_name	cell_line	pred
0	5-FU	DINACICLIB	A2058	4.538417
1	5-FU	DINACICLIB	A2780	3.973541
2	5-FU	DINACICLIB	A375	5.411381
3	5-FU	DINACICLIB	A427	1.086729
4	5-FU	DINACICLIB	CAOV3	0.801080

For data normalization, I first standardized & applied hyperbolic tangent, then standardized and applied hyperbolic tangent again, then (iii) standardized and applied the hyperbolic tangent again. The output layer was activated using linear activation, whereas the hidden layers were activated using corrected linear activation (Nair and Hinton, 2010). The aim was to reduce the mean squared error. With first hidden layer's 2048, 4096, or 8192 neutrons, it was possible to have two or three hidden layers. We observed that rectangular layers contain the same amount of neurons in the hidden layer as conic layers, but conic layers have half the amount of units. It was decided to apply early-stopping and dropout regularization

techniques, as well as SGD learning rates of 102 in our optimization. A moving average of 25 epochs on a validation set was used to determine early-stopping training iterations. Our dropout rates for input and the hidden layers were both 0.2, whereas our dropout rate was zero. Grid search was used to find the best hyper parameter.

3.1.9.4 Stratified Nested Cross-Validation

Comparing Deep Synergy's performance to other techniques, we employed stratified nested cross-validation. Various cross-validation approaches are shown in Figure 3. Random cross-validation, on the other hand, does not randomize the distribution of test samples across folds. We used a stratified cross-validation technique to exclude pharmaceutical combinations from the test populations (see Fig. 3 second column). The validation error in a 5-fold layered cross-validation setup guided the selection of the hyperparameters in an inner loop (Baumann and Baumann, 2014). Hyperparameter selection was then used to assess the inner loop model's performance on the outer test fold. For novel drugs (see Fig. 3 third column) as well as new cell lines, we employed stratified cross-validation to evaluate the generalizability of the processes.

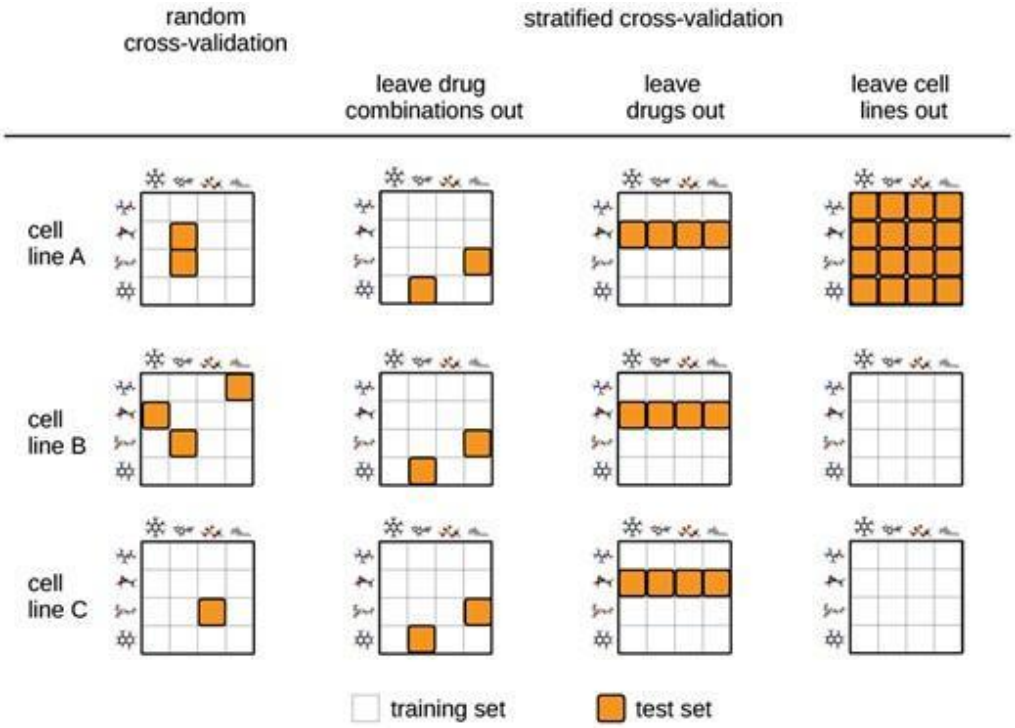


Figure 4

3.1.10 Results

Figure 4 depicts the Loewe model-based density distribution of synergy scores. 326 to 179 are the values that fall inside the range. The distribution's median is 4.37, and its standard deviation is 22.89. As a rule of thumb, everything with a synergy score greater than 0 is automatically considered synergistic. Clinical studies are interested in drug combinations that have a substantial synergistic effect. As a consequence, our attention was drawn to the top 10% of the population. A combination was considered positive synergistic if it had a measured score of more than 30. Drug combinations that were antagonistic, additive, or low-synergistic were grouped into the negative class.

Deep Synergy's test MSE was 255, whereas Gradient Boosting Machines & Random Forests scored 275 and 308 points, respectively. While Support Vector Machines & Elastic Nets both had MSEs of 398, median polish had an MSE of 478, which was used as a baseline. 53 percent of the time, the best approach outperforms the rest of the pack. If the increases in mean squared errors were statically important, the Wilcoxon signed rank-sum test was used with all P-values 0.05. Due to its superior MSE, RMSE, & Pearson correlation scores, Deep Synergy stands head and shoulders above the rest of the pack. With the help of new drugs and cell

lines, we also examined how the system operated. The prediction performance of all approaches on new drugs and novel cell lines is much worse than on novel drug combinations. Medication susceptibility scores range from 414 to 500, whereas cell susceptibility scores range from 387 to 461). Since the baseline approach does not take into account chemical or cell line features, the best-performing strategy only improves by 16 to 17 percent. Training examples in terms of chemical compounds (23 instances) and cell lines are suggested to be the cause of poor prediction performance (33 examples). Predictive synergy models should benefit greatly from bigger synergy datasets that include more chemical substances and cell lines.

3.1.11 Performance

It is difficult to compare RMSE and MSE across various datasets since they are reliant on the dataset. We also give performance metrics that are used in classification tasks to better evaluate Deep Synergy's predicted performance and provide comparative measures:. Accuracy and specificity (TPR and TNR), as well as Cohen's Kappa and ROC AUC;

Deep Synergy Architecture

Deep Synergy's architecture was determined using the hyperparameter selection approach, the results of which have been published. First, a hyperbolic tangent, and then a second standardization were determined to be the most successful steps in Tanh normalization. Deep Synergy also has conic layers. The regularizing effect of conic layers may account for how effectively they work. With only a few parameters in the upper layers, a generalization of the chemical characteristics of the input compound combinations is necessary to cope with this problem.In addition, the initial layer of 8192 units performed better. The reduced learning rate (105) and regularization of dropouts were also necessary for learning high-performance networks. There are two hidden layers in Deep Synergy's conic architecture, the first of which has 8192 neurons while the second of which has a total of 4096. With a learning rate of 105 and dropouts in input of 0.2 and 0.5, it makes advantage of input normalization.

3.1.12 Applicability Domain

We also calculated Pearson correlation coefficients to see how Deep Synergy fared across cell lines and medicines. Figure 5 displays the findings for (a) pharmaceuticals and (b) cell lines

on the left and right sides, respectively. Correlation values ranged from 0.57 to 0.84 for the drugs examined. In contrast, 39 percent of drugs may be predicted using correlation coefficients greater than 0.7. On the left side of Figure 5, there seems to be no connection between targets and correlation. As a consequence, target-specific procedures cannot account for performance discrepancies. There is minimal influence on performance if there are many drugs targeting the same target. Correlation values ranged from 0.57 to 0.84 for the drugs examined. In contrast, 39 percent of the drugs had correlation coefficients of greater than 0.7, while five prescriptions have correlation coefficients of less than 0.6. The left side of Figure 5 does not show any clear association between objectives and correlation. As a consequence, target-specific procedures cannot account for performance discrepancies. There is minimal influence on performance if there are many drugs targeting the exact same end goal. The cell lines studied have correlation values ranging from 0.56 to 0.84. A correlation of 0.6 or less is seen in just two cell lines. More over half of the cell lines have a correlation greater than 0.7 that can be anticipated. Figure 5 shows that the connection does not depend on the kind of tissue. Differences in performance are not explained by the quantity of cell lines derived from a particular tissue type.

3.1.13 Discussion

Our team created Deep Synergy, a revolutionary Deep Learning-based approach that allows us to accurately assess the synergy ratings of pharmaceutical combinations for cancer cell lines. This technique will only work if the cancer cell lines can be correctly identified using genetic and chemical descriptors. Otherwise, it will fail. Cross-validation using external test sets revealed that Deep Synergy outperformed all other techniques by a significant margin.Based on Deep Synergy predictions with an AUC of 0.90, the time and cost associated with the experimental validation of drug combinations may be reduced (Simm et al., 2018). Because of a scarcity of medications and cell lines in the dataset, generalization to novel therapies and cell lines is difficult to do. We are certain that this impediment will be overcome shortly, owing to the significant increase in dataset sizes over the preceding few years, as well as our expectation that this growth will continue in the future.Antifungals & antibiotics, both which make use of pharmacological combinations in their therapy, may benefit from our approach. Increased prediction performance for Deep Synergy may be achieved as a consequence of increasing the dataset size and introducing algorithmic enhancements,

respectively (Klambauer et al., 2017). Deep Synergy, we believe, may be used to uncover new synergistic medicine combinations that are effective.

References

Bates, S. (2010). Progress towards personalized medicine. *Drug Discovery Today*, *15*(3–4), 115–120.

Chiu, Y.-C., Chen, H.-I. H., Gorthi, A., Mostavi, M., Zheng, S., Huang, Y., & Chen, Y. (2020). Deep learning of pharmacogenomics resources: Moving towards precision oncology. *Briefings in Bioinformatics*, *21*(6), 2066–2083.

Delhalle, S., Bode, S. F., Balling, R., Ollert, M., & He, F. Q. (2018). A roadmap towards personalized immunology. *NPJ Systems Biology and Applications*, *4*(1), 1–14.

La Thangue, N. B., & Kerr, D. J. (2011). Predictive biomarkers: A paradigm shift towards personalized cancer medicine. *Nature Reviews Clinical Oncology*, *8*(10), 587–596.

Nevins, J. R., Huang, E. S., Dressman, H., Pittman, J., Huang, A. T., & West, M. (2003). Towards integrated clinico-genomic models for personalized medicine: Combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics*, *12*(suppl_2), R153–R157.

Osman, A. F. (2019). A multi-parametric MRI-based radiomics signature and a practical ML model for stratifying glioblastoma patients based on survival toward precision oncology. *Frontiers in Computational Neuroscience*, 58.

Parimbelli, E., Marini, S., Sacchi, L., & Bellazzi, R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, *83*, 87–96.

Piñeiro-Pérez, R., Abal, M., & Muinelo-Romay, L. (2022). Liquid Biopsy for Monitoring EC Patients: Towards Personalized Treatment. *Cancers*, *14*(6), 1405.

Pinker, K., Chin, J., Melsaether, A. N., Morris, E. A., & Moy, L. (2018). Precision medicine and radiogenomics in breast cancer: New approaches toward diagnosis and treatment. *Radiology*, *287*(3), 732–747.

Prados, M. D., Byron, S. A., Tran, N. L., Phillips, J. J., Molinaro, A. M., Ligon, K. L., Wen, P. Y., Kuhn, J. G., Mellinghoff, I. K., & De Groot, J. F. (2015). Toward precision medicine in glioblastoma: The promise and the challenges. *Neuro-Oncology*, *17*(8), 1051–1063.

Regel, I., Mayerle, J., & Ujjwal Mukund, M. (2020). Current strategies and future perspectives for precision medicine in pancreatic cancer. *Cancers*, *12*(4), 1024.

Rello, J., Van Engelen, T. S. R., Alp, E., Calandra, T., Cattoir, V., Kern, W. V., Netea, M. G., Nseir, S., Opal, S. M., & van de Veerdonk, F. L. (2018). Towards precision medicine in sepsis: A position paper from the European Society of Clinical Microbiology and Infectious Diseases. *Clinical Microbiology and Infection*, *24*(12), 1264–1272.

Seyhan, A. A., & Carini, C. (2019). Are innovation and new technologies in precision medicine paving a new era in patients centric care? *Journal of Translational Medicine*, *17*(1), 1–28.

Torres, C., & Grippo, P. J. (2018). Pancreatic cancer subtypes: A roadmap for precision medicine. *Annals of Medicine*, *50*(4), 277–287.

Vargas, A. J., & Harris, C. C. (2016). Biomarker development in the precision medicine era: Lung cancer as a case study. *Nature Reviews Cancer*, *16*(8), 525–537.

West, M., Ginsburg, G. S., Huang, A. T., & Nevins, J. R. (2006). Embracing the complexity of genomic data for personalized medicine. *Genome Research*, *16*(5), 559–566.