

Article

A Novel Recurrent Neural Network Framework for Prediction and Treatment of Oncogenic Mutation Progression

Rishab Parthasarathy ^{1,*}  and Achintya K. Bhowmik ² 
¹ Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

² Department of Otolaryngology-Head and Neck Surgery, Stanford University School of Medicine, Stanford, CA 94305, USA; akbhowmik@stanford.edu

* Correspondence: rpartha@mit.edu

Abstract

Despite significant medical advancements, cancer remains the second leading cause of death in the US, causing over 600,000 deaths per year. One emerging field, pathway analysis, is promising but still relies on manually derived wet lab data, which is time-consuming to acquire. This work proposes an efficient, effective, end-to-end framework for Artificial Intelligence (AI)-based pathway analysis that predicts both cancer severity and mutation progression in order to recommend possible treatments. The proposed technique involves a novel combination of time-series machine learning models and pathway analysis. First, mutation sequences were isolated from The Cancer Genome Atlas (TCGA) Database. Then, a novel preprocessing algorithm was used to filter key mutations by mutation frequency. This data was fed into a Recurrent Neural Network (RNN) that predicted cancer severity. The model probabilistically used the RNN predictions, information from the preprocessing algorithm, and multiple drug-target databases to predict future mutations and recommend possible treatments. This framework achieved robust results and Receiver Operating Characteristic (ROC) curves (a key statistical metric) with accuracies greater than 60%, similar to existing cancer diagnostics. In addition, preprocessing played a key role in isolating a few hundred key driver mutations per cancer stage, consistent with current research. Heatmaps based on predicted gene frequency were also generated, highlighting key mutations in each cancer. Overall, this work is the first to propose an efficient, cost-effective end-to-end framework for projecting cancer prognosis and providing possible treatments without relying on expensive, time-consuming wet lab work.

Keywords: recurrent neural networks; computational biology; time-series analysis



Academic Editors: Emanuele Frontoni and Miguel Angel Cazorla

Received: 30 November 2025

Revised: 10 January 2026

Accepted: 14 January 2026

Published: 2 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Cancer remains a major challenge worldwide, and despite numerous improvements in treatment over the years, is still the second leading cause of death in the United States, only behind heart disease, with over 600,000 deaths every year [1].

There are three main causes for the continuing challenges associated with treating cancer. First, complex, late-stage cancers are either often untreatable or exhibit resistance to treatments such as chemotherapy [2–4]. Second, at least 25% of cancer is not caught early, reducing effective treatment outcomes [5]. Third, when signs of precancerous progression are discovered, there is often no way to treat it without surgery [4]. Thus, new methods for early detection and treatment are crucial in saving lives.

Today, modern methods for early detection and treatment of cancer involves a three-step process. Starting with annual physical examinations, medical professionals determine any abnormalities in the patient's health. If any abnormalities are detected, patients are subjected to a series of scans and biopsies, which allow doctors to localize and identify any possible cancerous lesions. Using this knowledge, doctors evaluate both prognosis and progression in order to properly identify and treat disease outcomes. An example of this paradigm can be found in Figure 1, which depicts a simplified workflow of how a group of oncologists diagnosed various cases of thyroid cancer [6].

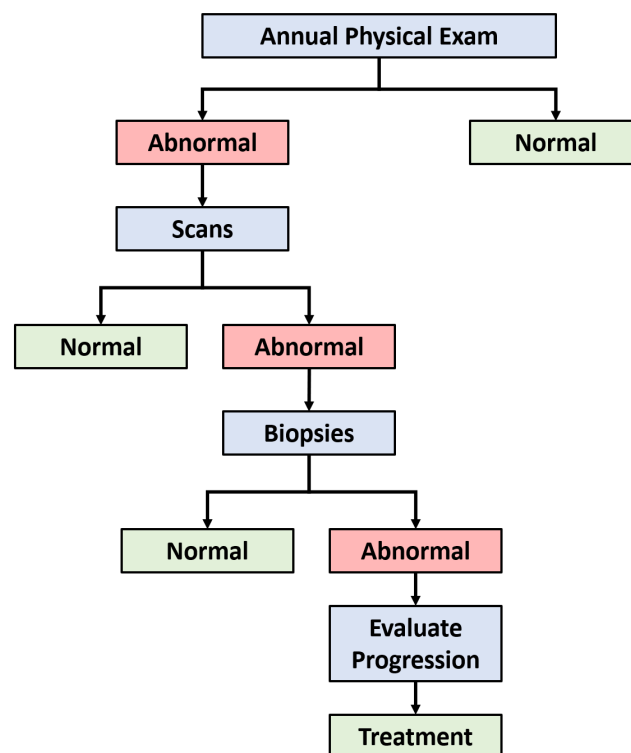


Figure 1. A sample, simplified flow chart that breaks down how oncologists diagnosed cases of thyroid cancer [6]. This image is reproduced with permission from Springer Nature with small stylistic modifications from [6].

Despite the effectiveness of this medical approach, no fully automated algorithm for evaluating this end-to-end pipeline exists, as current computational approaches can only analyze scans and biopsies at a fixed point in time [7].

1.1. Prior Research

1.1.1. Biological Research

Recent work in the field of bioinformatics has focused on investigating the use of computational models for analyzing patient scans and biopsies [8]. Many approaches have been developed which diagnose cancer from images of Magnetic Resonance Imaging (MRI) scans, primarily using Convolutional Neural Networks (CNNs) [8–10]. Further advancements have focused on the use of segmentation algorithms, which build on CNN-based techniques by additionally isolating the cancerous region detected in the scan [11–17].

However, while these methods are effective at diagnosing cancer from single images, they are not able to provide an analysis of a patient's disease progression. Thus, as gene sequencing has become increasingly accessible, recent research has moved towards automated genomic analyses using Deep Neural Networks (DNNs) [18,19]. In DNN-based genomic analysis, genomic data is fed into a series of Fully Connected Layers, which

calculate correlations between all pairs of genes [20,21]. These correlations can then be used to construct a more complete picture of a patient's cancer progression and prognosis.

In addition to computational methods, researchers have tackled the genomic aspects of cancer by developing target drugs and gene therapy. Target drugs function by inhibiting specific genes that are crucial to a given cancer's behavior, limiting progression and improving prognosis [22]. For example, target drugs have recently been used to treat previously intractable late-stage renal cancers [23]. On the other hand, gene therapy replaces or inactivates faulty genes in order to revert cancer cell development [24].

However, both these approaches are significantly limited by the overhead of finding the correct gene to target [25]. While both gene therapy and target drugs are effective when the correct gene is discovered, the incorrect gene is often targeted, resulting in ineffective treatment [25,26]. This key challenge, in combination with the high cost of gene-based treatments, has made gene therapy and target drugs difficult to scale [25].

One possible solution to this challenge is an emerging approach called pathway analysis, which combines both gene–drug and gene expression correlations into one biological interpretation [27]. Pathway analysis involves the calculation of coefficients regarding gene interaction or expression, which reveal biological correlations termed “pathways” [28]. These pathways have already proved successful in discovering gene–drug combinations for therapeutic purposes [29–31]. For example, a simplified snapshot of a pathway analysis framework for Head and Neck Squamous Cell Carcinoma (HNSCC) is presented in Figure 2, which depicts gene–gene interactions as lines between circles and gene–drug interactions as lines between yellow circles and red squares [32].

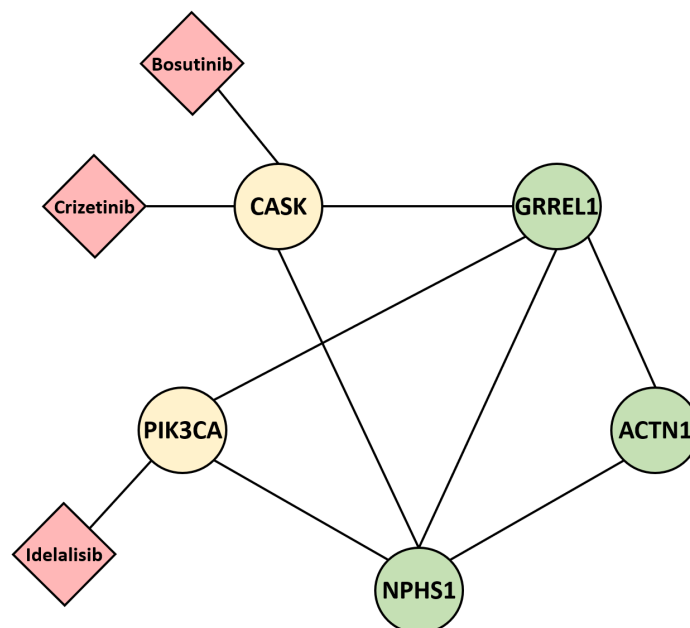


Figure 2. A simplified sample of a snapshot of discovered biological pathways based on manual computation of gene expression in Head and Neck Squamous Cell Carcinoma (HNSCC). Gene–gene interactions are depicted as lines between circles and gene–drug interactions as lines between yellow circles and red squares. These pathways have to be calculated and evaluated by hand for verification purposes [32].

However, pathway analysis is limited because it depends on manual processing of wet lab RNA-seq data, which is a time-consuming process [27,29]. Specifically, since each biological relationship in a pathway may require a long period of time to occur, observing each interaction in the pathway may take significant periods of time in the

lab. Hence, computational solutions may be able to limit this burden by approximating biological pathways without the need for wet-lab analysis. Thus, by integrating biological pathways with time-series analysis models, this paper aims to derive a new computational methodology for approximating biological pathways through time.

1.1.2. Time-Series Analysis and Recurrent Neural Networks (RNNs)

In recent years, neural network-based approaches for time-series analysis have become increasingly prominent, with many successful applications in fields such as natural language processing. One key framework for neural-network based time-series analysis is the Recurrent Neural Network (RNN), which has been used for a range of tasks from generating Shakespearean plays and language to time-series analyses of the stock market [33,34]. In these applications, RNN-based machine learning architectures have proved successful because of their ability to comprehensively generate correlations through time and order [33–35].

Specifically, the RNN architecture utilizes the idea of “attention” [35]. Attention is an algorithm where existing results from previous time-steps are amplified in order to make more informed decisions at the current time-step, essentially implementing a human-like understanding of past context [36]. For example, in a language-based model, given the sentence, “The archer wields a bow,” an attention-based model would be able to understand that the word “bow” means the archer’s weapon, not the act of bowing.

Similar to language, mutation sequences often exhibit correlations through time, but despite this apparent connection, RNNs have not yet been comprehensively used for evaluating the progression of cancer mutations [37]. This project also elects to use RNNs over Transformers, another leading attention-based framework, because Transformers require a significantly larger dataset to train effectively, only serving to increase the model’s complexity and run-time [36]. Thus, this project attempts to investigate the parallel between time-series and genomic data by employing RNNs for genomic analyses.

1.1.3. Deep Learning for Genomics

This work is further supported by recent advancements in deep learning for genomics, which have highlighted the strength of deep learning-based solutions for genomic sequence modeling. Recent work has concluded that sequential models like Transformers possess sufficient granularity to predict DNA sequences at the single-nucleotide level, generating entire patient DNA sequences from scratch [38–40]. However, while these works demonstrate the relevance of sequential deep learning models for genomic modeling, recent work has primarily focused on foundation models for DNA sequence generation as a whole. Instead, this work focuses on using sequential deep learning models for targeted treatment and prediction of cancer progression.

Outside of genomic sequence modeling, recent work in deep learning for genomics has focused on integrative multi-omics approaches, which use multiple data streams to create more informative predictions. These multi-omics approaches have integrated multiple data streams from genetic data to image-based data to predict cancer prognosis. However, these approaches largely use deep learning-based predictions as inputs for existing prognosis and gene-based models, specifically the Cox proportional hazards model [41–44]. On the other hand, existing RNN-based genomic analyses do not predict prognosis, and instead only predict mutational load as a proxy for driver mutations [45].

In terms of end-to-end analysis for predicting prognosis, existing approaches still utilize non-sequential models, so this work serves as a proof-of-concept for sequential RNN-based approaches for predicting prognosis. Due to limited data, this work focuses on stage-based analysis, which is reflected in other recent works as well [46–49].

1.2. Objectives

Hence, to better model how doctors diagnose patients, new models for cancer diagnostics must evaluate and treat possible disease progression. Recent advances in genomics and Artificial Intelligence (AI) have provided significant opportunities for developing a complete cancer diagnostic framework that can provide more systematic aid to patients.

This work draws on recent advances in machine learning-based time-series analysis, specifically Recurrent Neural Networks (RNNs) [33,34]. This work strives to apply the same sequential paradigm to cancer mutation sequences, extracting contextual information from each mutation. In doing so, this work aims to predict not only the present state of cancer, but also the future progression of the disease, possibly unveiling ways to treat cancer symptoms before they even occur. This overall methodology, based on RNN models consisting of Long Short-Term Memory (LSTM) architectures, is portrayed in Figure 3.

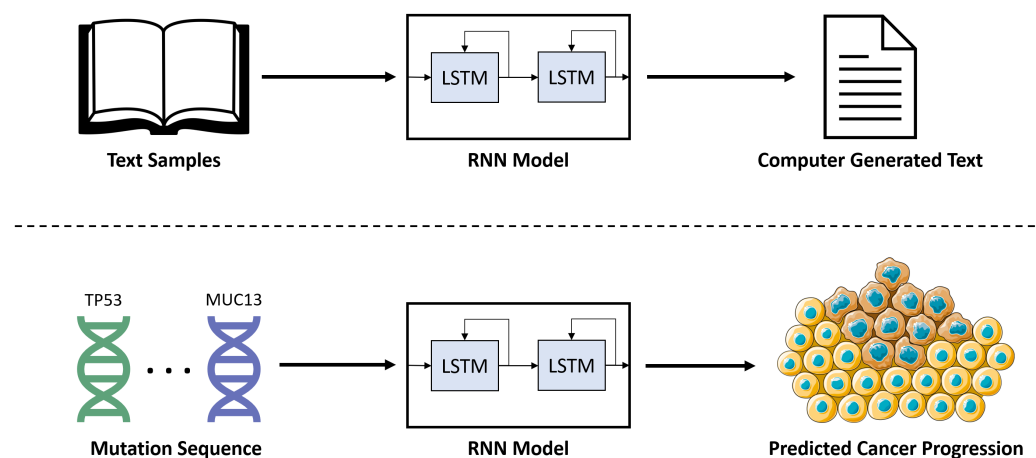


Figure 3. An illustration depicting the parallels between the processing of language and this project's methodology of approaching genomics. In both cases, the input data of text or mutations are fed into an RNN, which learns to infer what will happen in the future through developing spatial and temporal correlations.

Overall, this work presents an efficient and effective end-to-end framework for machine analysis of biological pathways to predict and prevent cancer progression. Using a novel RNN-inspired approach to pathway analysis, this framework provides functionalities for diagnosing cancer, evaluating cancer prognosis, and developing post hoc progression prediction and treatment recommendations using genomic data from a patient's tumor. The goal of this research is to help reduce the burden of cancer on hospitals, doctors, and patients by producing a methodology for targeted treatment of future genomic mutations, demonstrating the feasibility of creating a comprehensive solution for cancer diagnostics.

2. Materials and Methods

2.1. End-to-End Framework

In this work, a novel methodology for comprehensive analysis of cancer prognosis and progression was developed based on the use of genomic information. As depicted in Figure 4, there are three phases to the methodology: (1) Data Processing, (2) Network Module, and (3) Result Processing.

In the Data Processing phase, a preprocessing algorithm was developed to extract the salient information from The Cancer Genome Atlas (TCGA) dataset, filtering for the most common mutations per stage [50]. After the data was filtered, the RNN network was trained. Once the model was trained, the RNN predicted patient prognosis, which was

used in combination with information from the preprocessing algorithm to predict disease progression and recommend drugs.

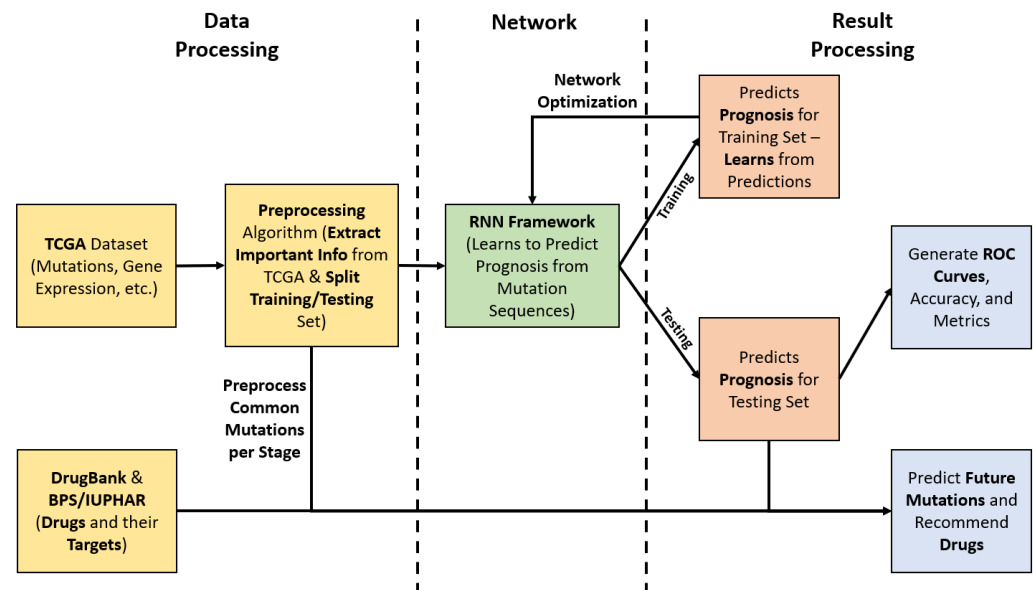


Figure 4. An illustration of the full end-to-end methodology. The Cancer Genome Atlas (TCGA) Dataset was preprocessed in order to find the most salient mutations and split the training/testing set. Then, the RNN framework was trained on the training dataset to accurately predict prognosis. The performance of the RNN was evaluated on the testing dataset, generating stage predictions which were used to generate accuracy and Receiver Operating Characteristic (ROC) curves. Finally, the predicted stages, the preprocessed list of important mutations, and the drug databases were used to predict future mutations and drug recommendations.

2.2. Dataset

In this work, three different datasets were used. The first dataset used was the TCGA dataset, which is the largest open-source genomic dataset for cancer, containing more than 20,000 patient mutation sequences. For this project, the TCGA dataset was extracted from cBioPortal, an online data repository for cancer genomics [51,52]. The TCGA dataset contains a detailed list of somatic mutations for each patient, along with a summary of the patient's type and severity of cancer [50]. Whenever possible, multiple timepoints for each patient were used; otherwise, cancer stage was used to generate a time-series, as cancer stage represents the progression of cancer through time. This progression using stages has been biologically validated, as different cancer stages correlate with distinct changes in cancer metabolism as metastasis develops [53–58]. Within each generated time point, mutations were ordered by chromosome. Within each chromosome, mutations were ordered by the start position of the mutation nucleotides.

After being extracted from cBioPortal, the classes in the TCGA dataset were evaluated for robustness in training and testing. A hard cutoff of at least 300 samples per class was set, and classes without genomic mutation data were eliminated. As a result, the TCGA dataset used consisted of 11 classes: Bladder Carcinoma (BLCA); Breast Carcinoma (BRCA); Colon Adenocarcinoma (COAD); Head-Neck Squamous Cell Carcinoma (HNSC); Kidney Renal Clear Cell Carcinoma (KIRC); Liver Hepatocellular Carcinoma (LIHC); Lung Adenocarcinoma (LUAD); Lung Squamous Cell Carcinoma (LUSC); Skin Cutaneous Melanoma (SKCM); Stomach Adenocarcinoma (STAD); and Thyroid Carcinoma (THCA) [50].

The other two datasets in this project were both used for drug discovery purposes, leveraging existing knowledge of drug–target correlations in order to provide targeted treatment plans. DrugBank, an open-source database run by the University of Alberta,

provided the bulk of drug–gene relationships [59–63]. In order to ensure the safety and efficacy of the drug treatments discovered, the International Union of Basic and Clinical Pharmacology/British Pharmacological Society (IUPHAR/BPS) Guide to Pharmacology database was used to validate the data in the DrugBank database [64].

2.3. Data Preprocessing

To make the TCGA dataset compatible with the RNN framework, a number of pre-processing techniques were applied, which was crucial because of two main challenges. First, many mutations were too rare to have verifiable impacts: for example, in the TCGA BRCA data, only 16.8% of mutations occurred in more than 1% of patients (10 patients in total) [50]. Second, the most expressed mutations were often the most clinically significant: clinical research had already verified that frequently observed mutations such as PIK3CA, TP53, and BRCA1 were key driver mutations in some of the most aggressive, lethal cancers [65,66].

To preprocess, the algorithm determined the most frequently expressed mutations both overall and in each stage. Based on the expression rates, the algorithm combined the mutation expression list from each stage, creating a list of significant mutations. The algorithm then filtered the TCGA input data to only contain such mutations. Once the data was filtered, the algorithm balanced the class sizes to prevent model overfitting. All in all, this preprocessing method not only simplified the network’s task but also caused increases in the performance as well. The entire preprocessing paradigm is presented in Figure 5.

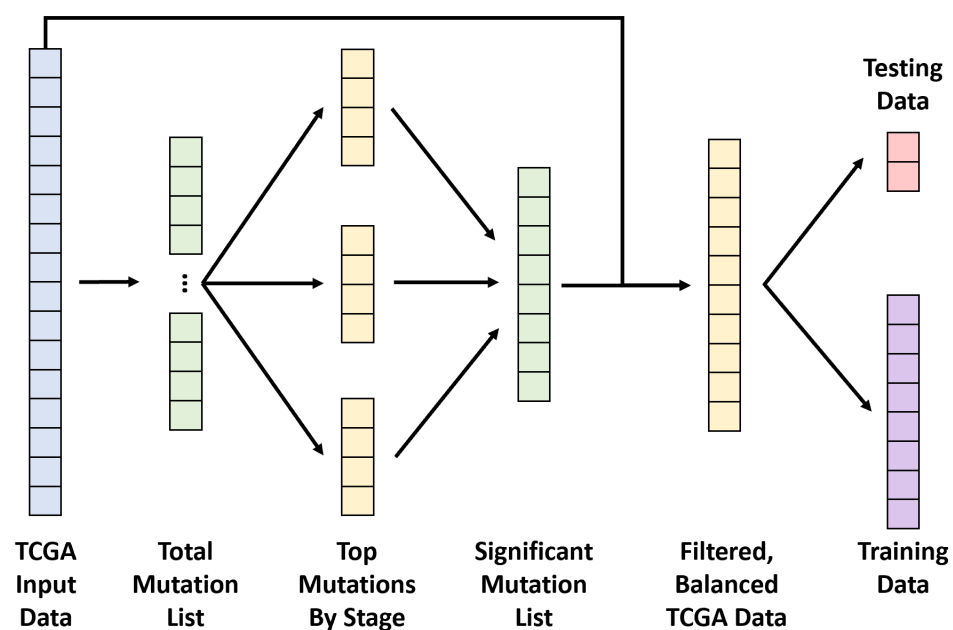


Figure 5. The entire preprocessing paradigm, from filtering to balancing the class size. In the first stage of preprocessing, a total mutation list was created used the TCGA input data before we calculated the most common mutations both by stage and overall. Then, these commonly observed mutations were used to create a significant mutation list, which was used to filter and balance the TCGA data, which was finally split into a training/testing split.

Specifically, each stage which constituted less than 10% of total data was removed from the data, as there was not enough data to be statistically significant. This modification also helped combat the rapid rate of overfitting that is inherent to deep neural networks [67].

Then, S_x was calculated as the top x mutations overall, and $S_{x,y}$ was calculated as the top x mutations in stage y , sorted by the expression frequency. Using these computed sets, the full mutation list was calculated using Equation (1).

$$S = \left\{ S_x, \dots, \left(S_{x,i} - \left(S_{x,i} \cap \left(S_x \cup \left(\bigcup_{j=1}^{i-1} S_{x,j} \right) \right) \right) \right) \right\} \quad (1)$$

Once S was computed, the preprocessing algorithm removed all mutations that were not selected from the dataset. We note that these mutation sets were computed using the full dataset, which does not contribute to data leakage as, no matter what, the model sees the same mutations at both train and test time.

Ultimately, the dataset was balanced by defining a weighted SoftMax transform, depicted in Equation (2), where for a sample vector v and weight vector w , the output P was calculated by [68]:

$$P_i = \frac{e^{v_i w_i}}{\sum_j e^{v_j w_j}} \quad (2)$$

To optimize the weighting, as shown in Equation (3), the weight vector w was defined using the class sizes c from the data, where

$$w_i = \frac{\sum_i c_i}{2c_i} \quad (3)$$

This weighting method prevented overfitting by equalizing the gradients created by each class within the training procedure.

2.4. Recurrent Neural Network (RNN)

The RNN framework used in this project followed a three-step model that used a sequence of text to generate predictions. In this case, each patient's mutation sequence was used to predict the cancer stage and generate temporal correlations between mutations.

The first step of the RNN was one-hot embedding, which signified that each mutation was processed as an array of all zeros apart from a single one. The embedding layer then transformed this mutation array into a shorter array of k bounded values. Specifically, this project utilized an embedding of length 256. The mathematical formalism for transforming a one-hot vector v of length n to an embedded vector e of length k is presented in Equation (4), given a matrix of weights w [69].

$$e_j = \sum_{i=1}^n v_i w_{ij} \quad (4)$$

By training the weights, the embedding learned correlations through the similarity between the embedded values.

The second step of the RNN was a series of Long Short-Term Memory (LSTM) units, which obtained one more piece of information for each time-step (each mutation read) [35]. The LSTMs could then learn temporal correlations in the data, which enabled the prediction of cancer progression.

This project employed a bidirectional LSTM layer, which simultaneously processed the data in both backward and forward directions. The forward pass trained the algorithm while the backward pass smoothed the predictions, allowing more data to be accurately analyzed [35,70]. A bidirectional LSTM layer is presented in Figure 6.

After the LSTM layer, the third and final step of the RNN was a series of fully connected or Dense layers, where each pair of sequential neurons was connected, enabling easy consolidation of information [71].

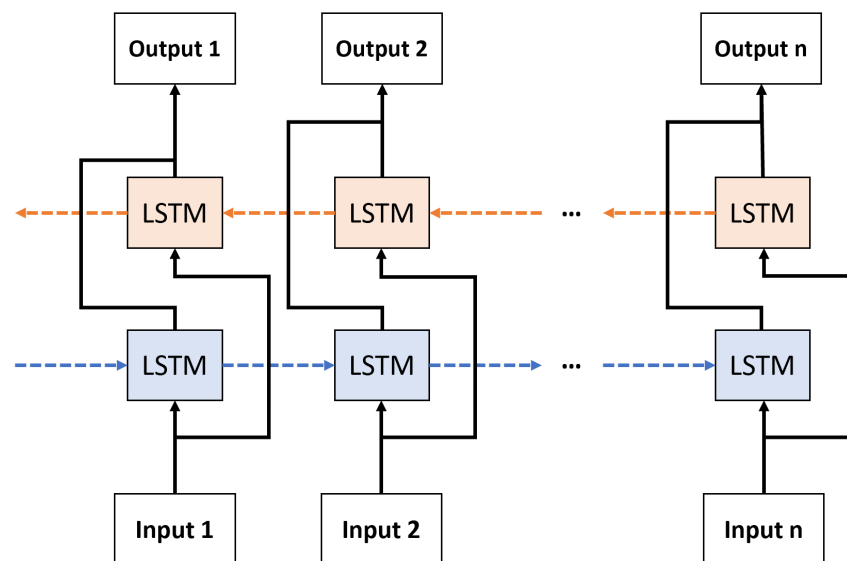


Figure 6. A sample bidirectional LSTM layer, where the blue represents the forward training pass of the algorithm, the orange represents the backward smoothing pass of the algorithm, and the black arrows represent the flow of data through the neural network.

2.5. Experimental Setup and Implementation

Overall, the specific RNN machine learning configuration used in this project contained an Embedding of length 256 (i.e., transforming each mutation into a float matrix of length 256), a bidirectional LSTM layer of length 64, and two dense layers, which were activated with the Rectified Linear Unit (ReLU) and SoftMax, respectively. A breakdown of this network is presented in Figure 7.

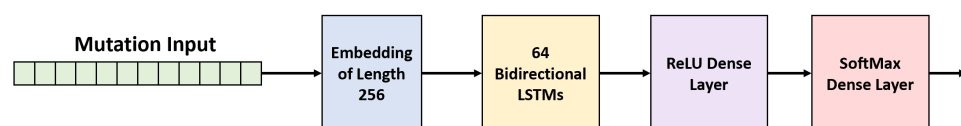


Figure 7. A breakdown of the network structure used, from the Embedding of length 256 and bidirectional LSTM layer of length 64 to the two dense layers, which were activated with the Rectified Linear Unit (ReLU) and SoftMax, respectively.

Training was performed using the Adam optimizer with a learning rate of 1×10^{-4} over 200 epochs and a batch size of 16. Sequences were postfix-padded to the maximum length in the dataset using empty strings. No early stopping strategies were used, and convergence was analyzed by continued decrease in validation loss. No random seeds were used. These parameters were generated using hyperparameter sweeps that checked for the highest validation accuracy.

2.6. Post Hoc Gene–Drug Prediction

Once the RNN framework produced a stage prediction, the algorithm then produced post hoc future gene predictions and generated drug treatments for those predicted genes.

First, using the RNN prediction, the algorithm extracted the mutations correlated with the predicted stage. These mutations were compared against the input mutation list to extract the mutations that had not yet occurred.

After extracting these significant future mutations, the postprocessing algorithm calculated the probability of each mutation occurring. This probability was extracted by evaluating the frequency at which each future mutation occurred relative to each stage.

In addition, with the driver mutation lists for cancer progression, the algorithm queried the DrugBank and IUPHAR/BPS databases of drug–target interactions, which described how certain drugs modified the behavior of given genes [59–64]. Using this information, the algorithm evaluated whether any treatments would treat predicted driver mutations, validating the DrugBank data using the IUPHAR/BPS database. A depiction of this pipeline is provided in Figure 8.

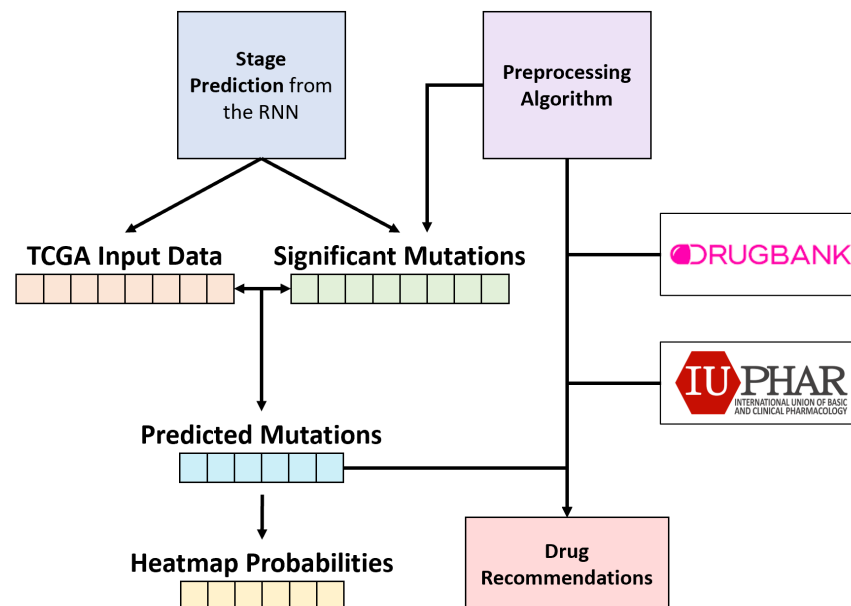


Figure 8. The pipeline used for gene–drug prediction, extracting both heatmaps of significant mutations and providing drug recommendations to treat these mutations even years into the future.

3. Results

Each model was trained for 200 epochs, with 80% of the data assigned to the training set and 20% assigned to the testing set. Various degrees of preprocessing were utilized in order to validate the effectiveness of the preprocessing algorithm. Specifically, preprocessing for the top 50, 100, and 200 mutations was tested for each cancer.

When relevant, algorithmic performance was evaluated using Receiver Operating Characteristic (ROC) curves, which plot sensitivity against specificity. The performance of ROC curves can be qualitatively evaluated by comparing the curves against the diagonal line (random guessing): a consistent lack of intersection between the curves and the line indicates robustness in the information that the algorithm learned [72]. Because ROC curves plot sensitivity against specificity, they provide a robust, strong test of predictive power and statistical significance compared to a random predictor [73]. The ROC curves were generated by running the final trained RNN on the test set, comparing the predicted stage to the actual stage.

3.1. Stage Predictions

To evaluate the effectiveness of the RNN algorithm in predicting cancer stage, the algorithm was run individually on the dataset from each cancer type, and both ROC curves and accuracy were generated. One ROC curve was generated for each cancer stage, and they were grouped by cancer type as presented in Figure 9.

For the purpose of clarity, Figure 9 presents a representative sample of the cancer types tested, distributed throughout different sections of the body. Thyroid cancer represents the endocrine system, kidney cancer represents the excretory system, head/neck cancer represents the nervous system, and breast cancer represents the lymphatic system.

These results demonstrate that all four models are robust, with ROC curves significantly above the diagonal. In addition, given that no individual ROC curve intersects with the diagonal, the model did not overfit at any specific stage. This behavior confirms the efficacy of the stage weighting procedure used during the preprocessing stage.

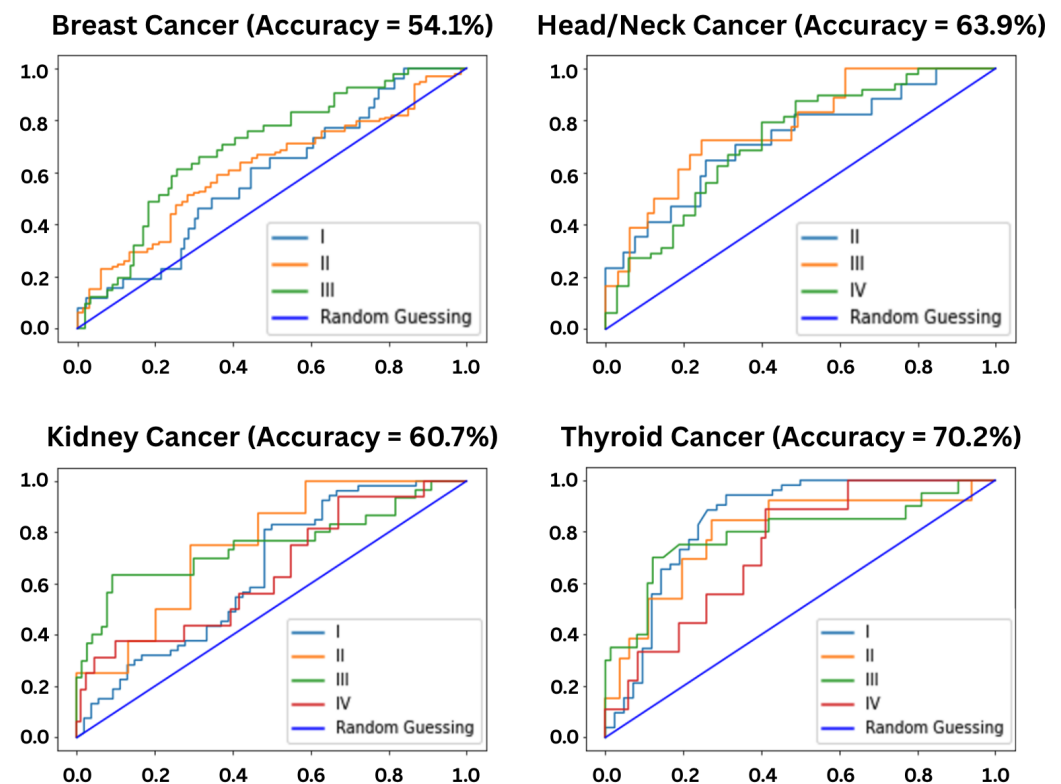


Figure 9. Receiver Operating Characteristic (ROC) curves from four of the cancers that this project evaluated (breast, head/neck, kidney, thyroid). Each ROC curve refers to the cancer stage that was predicted. These ROC curves are all robust, significantly above Random Guessing, which implies the model's successful retention of genomic attributes correlated with stage/severity.

3.2. Preprocessing Performance

The ROC curves also demonstrate important insights from preprocessing, as depicted in the representative examples provided in Figure 10. These results clearly indicate that the preprocessing methods enhanced the model's performance, improving the results from random guessing to true robust predictions, as in the case of breast cancer, with a 1.6-fold increase in accuracy: from 33.9% to 54.1%. In addition, the ROC curves demonstrate that eliminating non-driver mutations improved algorithmic performance significantly, indicating that the algorithm may not have been able to find long-term correlations from many mutations. However, as with head and neck cancer, preprocessing the top 200 expressed mutations yielded far better results than just 50 mutations, with a 1.75-fold increase between 63.9% and 36.6%. This massive increase in accuracy and robustness suggests that there may be around 200 key mutations in head/neck cancer, as there may not have been sufficient information for the model to learn from just 50 mutations. All other cancer types also had optimal performance when preprocessing for 200 mutations, compared to only 50 mutations or the whole dataset of thousands of mutations. This result may imply that the number of key mutations is on the order of a few hundred for the types of cancer analyzed in this study.

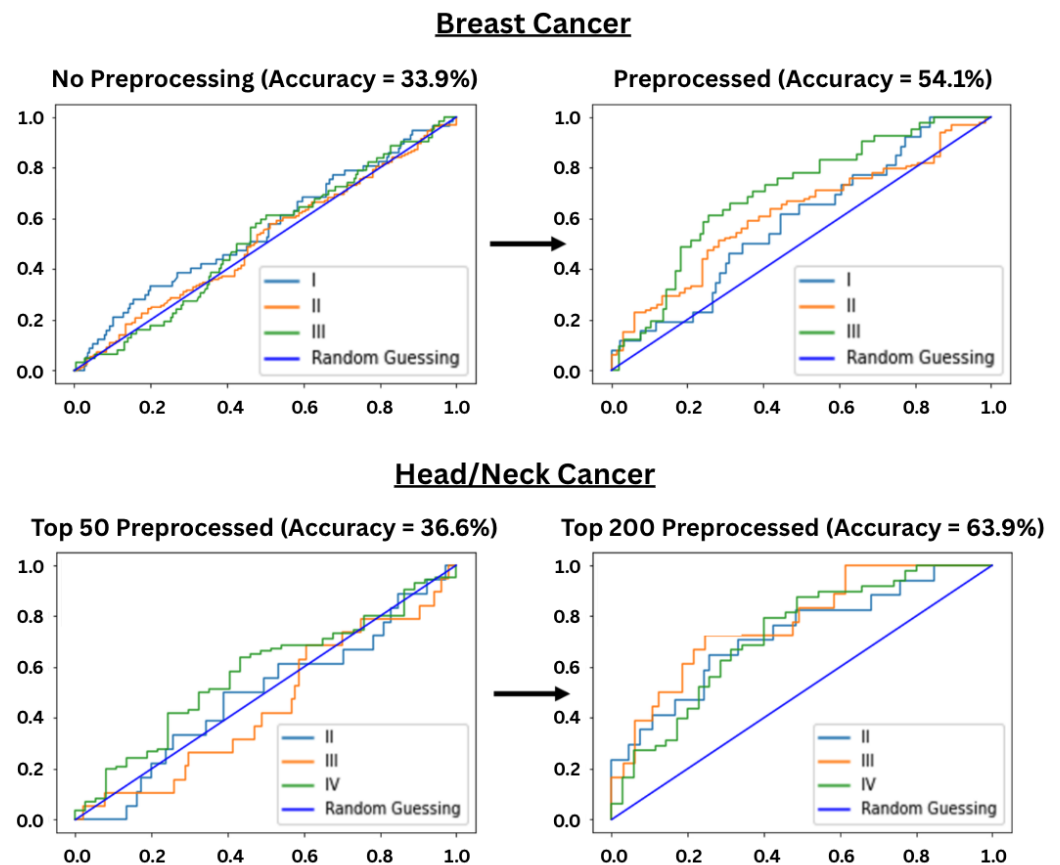


Figure 10. This figure depicts two insights from preprocessing using representative cancer types. First, preprocessing improves the algorithm performance, as many non-driver mutations are removed, as depicted with breast cancer. Second, preprocessing the top 200 most expressed mutations is most effective for robustness, indicating that there may be on the order of 200 key driver mutations.

4. Discussion

The framework presented in this paper is capable of computationally staging cancer progression using RNNs. This RNN-based framework has several key advantages over current genomic models. First, by learning from raw data, the model does not require humans to manually parse the input data to process sequential pathways. In essence, the model functions without relying on wet-lab RNA-sequencing data, which is time-consuming to produce [27,29]. Secondly, this model serves as a proof-of-concept that sequential models can be used for cancer prognosis prediction, instead of snapshot models that only process one step in time.

As for preprocessing, this investigation discovered that computational models are most effective when processing the top 200 mutations for each stage, which was observed over all 11 types of cancer investigated. There are two possible reasons for this observation. First, with a high number of mutations, rarely expressed mutations encouraged network overfitting, rendering performance inadequate on sequestered testing datasets. Second, utilization of smaller number of mutations decreased network robustness, implying that there may be key biological pathways specifically encoded within the order of a few hundred driver mutations. This result is consistent with other biological research that has discovered a similar order of a few hundred consistently observed genes with driver mutations [74,75].

The model's accuracy on genomic data then computationally verified a link between mutations and the cancer stage, demonstrating the predictive power of utilizing the temporal relationship between mutations. In addition, the prediction accuracy for stage (severity)

was either around or greater than 50–60%, which was comparable to both existing computational models and the performance of medical professionals in estimating cancer prognosis, as presented in Table 1, where GAN represents a Generative Adversarial Network, RF represents a Random Forest model, DNN represents a Deep Neural Network, GCN represents a Graph Convolutional Network, and BNN represents a Bayesian Neural Network [46–48,76–78].

Table 1. Comparative performance of different diagnosis frameworks.

	Model	Accuracy Range	Cancer Types
This Work	RNN	36–70%	11 types
López-García et al. [76]	CNN	68%	Lung
López-García et al. [76]	ML	62–70%	Lung
Kwon et al. [77]	GAN + CNN	41–80%	12 types
Kwon et al. [77]	GAN + RF	47–74%	12 types
Kwon et al. [77]	GAN + DNN	42–77%	12 types
Yu et al. [49]	C5.0	70–95%	8 types
Elmahy et al. [46]	GCN	82%	Renal
Gkatzamanidou et al. [47]	BNN	93%	Lung
Amanzholova et al. [48]	Ensemble	89–97%	3 types
Malhotra et al. [78]	Oncologists	62%	Advanced

Thus, while this model achieved comparable performance to a survey of oncologists and simple baselines, it lags behind the current state-of-the-art models, which achieve significantly higher accuracy, albeit on a much smaller array of cancer types. However, the fact that a simple RNN-based solution trained only on mutation sequences achieves similar performance to oncologists provides strong evidence that as a proof-of-concept, sequential models may have potential for prognostic prediction. Furthermore, Amanzholova et al. found that multi-omics data caused a twenty percent improvement in performance, indicating that future work on multi-omics sequential prediction may ultimately result in a useful diagnostic and prognostic aid for helping doctors project the progression of a patient’s disease. Hence, this work is still a useful proof-of-concept that simple RNN-based sequential modeling is comparable with other simple CNN/RF/DNN baselines, even if it lags behind more complicated state-of-the-art models, demonstrating the future potential of our work.

In terms of performance across cancer types, the one outlier in the model’s success was its performance on the Colorectal Adenocarcinoma (COADREAD) dataset, on which the model only achieved 36% accuracy even after preprocessing, as well as Lung Squamous Cell Carcinoma (LUSC) and Skin Cutaneous Melanoma (SKCM), where the model only achieved around 45% accuracy. Despite being competitive with the numbers proposed by Kwon et al., this may suggest one limitation of the model: that it cannot account for external factors such as lifestyle and environmental circumstances, which can play the most significant roles in causing cancers like melanoma (UV radiation), colorectal adenocarcinoma (diet), and lung squamous cell carcinoma (smoking) [77,79,80]. In addition, the TCGA dataset draws from a relatively limited pool of people, so further evaluation on larger, more equitable datasets will be necessary to truly scale this project [81].

Limitations of Stage-Based Prediction

However, one key limitation of the stage-based prediction framework is the lack of granularity in terms of prediction results. Specifically, the categorization of prognosis into four stages limits the possible estimations of prognosis, as different cancers of the same stage may have significantly different prognostic outcomes.

This work attempts to combat this using post hoc gene–drug analysis, where using the prediction from the RNN, existing gene and drug data can be used to predict possible future mutations and possible treatments.

For example, representing the pre-processed TCGA dataset using heatmaps, which depict the probability of observing each mutation in each stage, as in Figure 11, there are a few mutations that are strongly correlated with later stages of cancer, like CDH1 for breast cancer and CDKN2A for head and neck cancer [82–87]. Using the prediction from the model, if the stage was predicted to be late-stage, but these mutations were not present, the post hoc gene prediction pipeline, using the probabilities from the dataset, would strongly correlate these mutations with further cancer progression, providing targets for preventative treatment.

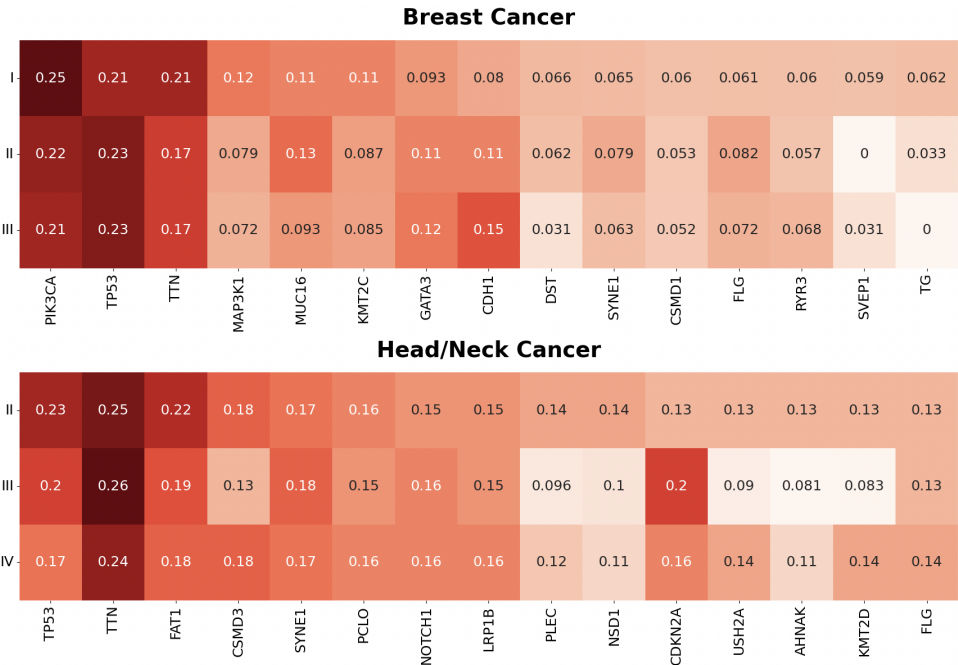


Figure 11. This figure presents two representative heatmaps of the TCGA dataset, correlating cancer stage on the vertical axis to individual cancer mutations on the horizontal axis. Each square represents the probability of finding a given mutation in a given stage, which is color coded with a darker red indicating a larger probability.

Similarly, with drug predictions, the algorithm can use the existing BPS/IUPHAR datasets to predict drugs based on the specific mutations provided. For example, if PIK3CA was indicated as a possible future mutation, the drug prediction pipeline would generate three possible treatments, alpelisib, copanlisib, and pilaralisib, which are all either in use as key FDA-approved treatments or in highly regarded clinical trials [88–96].

While limited, these post hoc methods provide techniques for scaling prognosis prediction to preventative treatment of future mutations. However, outside of time-series analysis, RNN-based models are also used for autoregressive generation of language sequences, which may also scale to autoregressive generation of mutation sequences through time, which should be investigated in future work.

5. Conclusions

Overall, this study was one of the first to apply AI frameworks based on RNN architectures, which are typically used for time-series analysis, to a genomic pathway analysis problem. By proposing, implementing, and evaluating an efficient, cost-effective end-to-end framework, this project demonstrates the proof-of-concept for an RNN-based model for predicting cancer severity and prognosis, which can be used for post hoc evaluation of progression and preventative treatments. In addition, by not relying on formally derived pathway correlations, this project enables rapid computational analysis of genomic data, allowing real-time prognosis prediction. In doing so, the model presented in this project may enable doctors to better analyze cancer prognosis, especially with additional improvements through adversarial training procedures [77].

This project has revealed the efficacy of applying a series-analysis-based approach to a genomic problem. In the future, analytical methods such as the use of Shapley values may be used to evaluate the internal RNN performance [97]. By unveiling the so-called “black-box” behind the RNN, a continuation of this research may understand the specific techniques and insights that the RNN uses to learn correlations. By combining these computational insights with the existing knowledge of biological pathways, this model may be able to deepen the fundamental understanding of the connection between various genes. Further work should also investigate the scalability of RNNs as autoregressive predictors of mutation progression, building on the post hoc pipeline in this work to create a fully RNN-based progression predictor.

In addition, the general paradigm proposed in this project can be extended to other diseases with a genomic correlation, such as cystic fibrosis or Alzheimer’s [98,99]. Thus, this project serves as a proof-of-concept for an efficient, cost-effective, and generalizable methodology for projecting disease prognosis, which in the future, could be life-saving, assisting with the prevention, diagnosis, and treatment of any genomically correlated disease, cancer and beyond.

Author Contributions: Conceptualization, R.P. and A.K.B.; methodology, R.P. and A.K.B.; software, R.P.; validation, R.P. and A.K.B.; formal analysis, R.P.; investigation, R.P.; resources, R.P. and A.K.B.; data curation, R.P.; writing—original draft preparation, R.P.; writing—review and editing, R.P. and A.K.B.; visualization, R.P.; supervision, A.K.B.; project administration, A.K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this project is publicly available at TCGA dataset <https://portal.gdc.cancer.gov/> (accessed on 17 January 2021), DrugBank database <https://go.drugbank.com/> (accessed on 17 January 2021), IUPHAR/BPS database <https://www.guidetopharmacology.org/> (accessed on 17 January 2021) [50–52,59–64]. The codebase developed and used can be found at <https://github.com/rishab-partha/Cancer-Progression-Pub> (accessed on 17 January 2021). All derived data is available upon reasonable request.

Acknowledgments: The authors would like to thank the teachers at the Harker School, especially Chris Spenner and Eric Nelson, for their contributions to reviewing early versions of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [\[CrossRef\]](#)
2. Housman, G.; Byler, S.; Heerboth, S.; Lapinska, K.; Longacre, M.; Synder, N.; Sarkar, S. Drug resistance in cancer: An overview. *Cancers* **2014**, *6*, 1769–1792. [\[CrossRef\]](#)
3. Riggio, A.I.; Varley, K.E.; Welm, A.L. The lingering mysteries of metastatic recurrence in breast cancer. *Br. J. Cancer* **2021**, *124*, 13–26. [\[CrossRef\]](#)
4. Rawla, P.; Sunkara, T.; Gaduputi, V. Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World J. Oncol.* **2019**, *10*, 10–27. [\[CrossRef\]](#)
5. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Tonorezos, E.S.; Barnea, D.; Moskowitz, C.S.; Chou, J.F.; Sklar, C.A.; Elkin, E.B.; Wong, R.J.; Li, D.; Tuttle, R.M.; Korenstein, D.; et al. Screening for thyroid cancer in survivors of childhood and young adult cancer treated with neck radiation. *J. Cancer Surviv.* **2017**, *11*, 302–308. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Xue, Y.; Wilcox, W.R. Changing paradigm of cancer therapy: Precision medicine by next-generation sequencing. *Cancer Biol. Med.* **2016**, *13*, 12–18. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Chougrad, H.; Zouaki, H.; Alheyane, O. Deep Convolutional Neural Networks for breast cancer screening. *Comput. Methods Programs Biomed.* **2018**, *157*, 19–30. [\[CrossRef\]](#)
9. Ha, R.; Chang, P.; Mema, E.; Mutasa, S.; Karcich, J.; Wynn, R.T.; Liu, M.Z.; Jambawalikar, S. Fully Automated Convolutional Neural Network Method for Quantification of Breast MRI Fibroglandular Tissue and Background Parenchymal Enhancement. *J. Digit. Imaging* **2019**, *32*, 141–147.
10. Jiang, Y.; Chen, L.; Zhang, H.; Xiao, X. Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. *PLoS ONE* **2019**, *14*, e0214587. [\[CrossRef\]](#)
11. Guo, Z.; Liu, H.; Ni, H.; Wang, X.; Su, M.; Guo, W.; Wang, K.; Jiang, T.; Qian, Y. A Fast and Refined Cancer Regions Segmentation Framework in Whole-slide Breast Pathological Images. *Sci. Rep.* **2019**, *9*, 882. [\[CrossRef\]](#)
12. Kurc, T.; Bakas, S.; Ren, X.; Bagari, A.; Momeni, A.; Huang, Y.; Zhang, L.; Kumar, A.; Thibault, M.; Qi, Q.; et al. Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches. *Front. Neurosci.* **2020**, *14*, 27. [\[CrossRef\]](#)
13. Mehta, S.; Mercan, E.; Bartlett, J.; Weave, D.; Elmore, J.G.; Shapiro, L. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Granada, Spain, 2018; pp. 893–901.
14. Işın, A.; Direkoğlu, C.; Şah, M. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. *Proc. Comput. Sci.* **2016**, *102*, 317–324.
15. Pereira, S.; Oliveira, A.; Alves, V.; Silva, C.A. On hierarchical brain tumor segmentation in MRI using fully convolutional neural networks: A preliminary study. In Proceedings of the 2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), Coimbra, Portugal, 16–18 February 2017; pp. 1–4.
16. Khan, M.K.H.; Guo, W.; Liu, J.; Dong, F.; Li, Z.; Patterson, T.A.; Hong, H. Machine learning and deep learning for brain tumor MRI image segmentation. *Exp. Biol. Med.* **2023**, *248*, 1974–1992. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Aslam, W.; Hussain, J.; Aslam, M.Z.; Jan, S.; Riaz, T.B.; Iqbal, A.; Arif, M.; Khan, I. Enhanced brain tumor segmentation in medical imaging using multi-modal multi-scale contextual aggregation and attention fusion. *Sci. Rep.* **2025**, *15*, 37308. [\[PubMed\]](#)
18. Collins, F.S.; Green, E.D.; Guttmacher, A.E.; Guyer, M.S. A vision for the future of genomics research. *Nature* **2003**, *422*, 835–847. [\[CrossRef\]](#)
19. Berger, M.F.; Mardis, E.R. The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 353–365. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Talukder, A.; Barham, C.; Li, X.; Hu, H. Interpretation of deep learning in genomics and epigenomics. *Brief. Bioinform.* **2021**, *22*, bbaa177.
21. Montesinos-López, O.A.; Montesinos-López, A.; Pérez-Rodríguez, P.; Barrón-López, J.A.; Martini, J.W.R.; Fajardo-Flores, S.B.; Gaytan-Lugo, L.S.; Santana-Mancilla, P.C.; Crossa, J. A review of deep learning applications for genomic selection. *BMC Genom.* **2021**, *22*, 19. [\[CrossRef\]](#)
22. Sawyers, C. Targeted cancer therapy. *Nature* **2004**, *432*, 294–297. [\[CrossRef\]](#)
23. Ghidini, M.; Petrelli, F.; Ghidini, A.; Tomasello, G.; Hahne, J.C.; Passalacqua, R.; Barni, S. Clinical development of mTor inhibitors for renal cancer. *Expert Opin. Investig. Drugs* **2017**, *26*, 1229–1237. [\[CrossRef\]](#)
24. Gonçalves, G.A.R.; Paiva, R.M.A. Gene therapy: Advances, challenges and perspectives. *Einstein* **2017**, *15*, 369–375. [\[CrossRef\]](#)
25. Buzdin, A.; Sorokin, M.; Garazha, A.; Sekacheva, M.; Kim, E.; Zhukov, N.; Wang, Y.; Li, X.; Kar, S.; Hartmann, C.; et al. Molecular pathway activation—New type of biomarkers for tumor morphology and personalized selection of target drugs. *Semin. Cancer Biol.* **2018**, *53*, 110–124. [\[CrossRef\]](#)

26. Gridelli, C.; De Marinis, F.; Di Maio, M.; Cortinovis, D.; Cappuzzo, F.; Mok, T. Gefitinib as first-line treatment for patients with advanced non-small-cell lung cancer with activating epidermal growth factor receptor mutation: Review of the evidence. *Lung Cancer* **2011**, *71*, 249–257. [CrossRef] [PubMed]
27. Khatri, P.; Sirota, M.; Butte, A.J. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **2012**, *8*, e1002375. [CrossRef] [PubMed]
28. Mziou-Sallami, M.; Roger, P.; Gloaguen, A.; Dandine-Roulland, C.; Ngaho, T.J.; Brohard, S.; Muret, K.; Sandron, F.; Bonnet, E.; Deleuze, J.-F.; et al. GNNenrich: A novel method for pathway enrichment analysis based on graph neural network. *Bioinformatics* **2025**, *41*, btaf478. [CrossRef]
29. Zolotovskaia, M.A.; Sorokin, M.I.; Emelianova, A.A.; Borisov, N.M.; Kuzmin, D.V.; Borger, P.; Garazha, A.V.; Buzdin, A.A. Pathway Based Analysis of Mutation Data Is Efficient for Scoring Target Cancer Drugs. *Front. Pharmacol.* **2019**, *10*, 1. [CrossRef] [PubMed]
30. Yang, X.; Kui, L.; Tang, M.; Li, D.; Wei, K.; Chen, W.; Miao, J.; Dong, Y. High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery. *Front. Genet.* **2020**, *11*, 19. [CrossRef]
31. Sivachenko, A.Y.; Yuryev, A. Pathway analysis software as a tool for drug target selection, prioritization and validation of drug mechanism. *Expert Opin. Ther. Targets* **2007**, *11*, 411–421. [CrossRef]
32. Choonoo, G.; Blucher, A.S.; Higgins, S.; Boardman, M.; Jeng, S.; Zheng, C.; Jacobs, J.; Anderson, A.; Chamberlin, S.; Evans, N.; et al. Illuminating biological pathways for drug targeting in head and neck squamous cell carcinoma. *PLoS ONE* **2019**, *14*, e0223639. [CrossRef] [PubMed]
33. Karpathy, A. The Unreasonable Effectiveness of Recurrent Neural Networks. 2015. Available online: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/> (accessed on 15 August 2022).
34. Moghar, A.; Hamiche, M. Stock Market Prediction Using LSTM Recurrent Neural Network. *Proc. Comput. Sci.* **2020**, *170*, 1168–1173. [CrossRef]
35. Shertinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
37. Zhu, W.; Xie, L.; Han, J.; Guo, X. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers* **2020**, *12*, 603. [CrossRef] [PubMed]
38. Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.W.; Sykes, C.B.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–19 December 2023.
39. Nguyen, E.; Poli, M.; Durrant, M.G.; Kang, B.; Katrekar, D.; Li, D.B.; Bartie, L.J.; Thomas, A.W.; King, S.H.; Bixi, G.; et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **2024**, *386*, eado9336. [CrossRef]
40. Lal, A.; Gunsalus, L.; Nair, S.; Biancalani, T.; Eraslan, G. gReLU: A comprehensive framework for DNA sequence modeling and design. *Nat. Methods* **2025**, *22*, 2253–2257. [CrossRef]
41. Hou, J.; Zhang, R.; Xie, Y.; Li, C.; Qin, W. Multimodal deep learning for cancer prognosis prediction with clinical information prompts integration. *npj Digit. Med.* **2025**, *9*, 76. [CrossRef] [PubMed]
42. Afreen, S.; Bhurjee, A.K.; Aziz, R.M. Cancer classification using RNA sequencing gene expression data based on Game Shapley local search embedded binary social ski-driver optimization algorithms. *Microchem. J.* **2024**, *205*, 111280. [CrossRef]
43. Niu, R.; Guo, Y.; Shang, X. GLIMS: A two-stage gradual-learning method for cancer genes prediction using multi-omics data and co-splicing network. *iScience* **2024**, *27*, 109387.
44. Nguyen, R.; Vafaee, F. Multi-omics prognostic marker discovery and survival modelling: A case study on multi-cancer survival analysis of women's specific tumours. *Sci. Rep.* **2025**, *15*, 36706. [CrossRef] [PubMed]
45. Auslander, N.; Wolf, Y.I.; Koonin, E.V. In silico learning of tumor evolution through mutational time series. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 9501–9510. [CrossRef]
46. Elmahy, A.; Aly, S.; Elkhowsky, F. Cancer Stage Prediction From Gene Expression Data Using Weighted Graph Convolution Network. In Proceedings of the 2021 2nd International Conference on Innovative and Creative Information Technology (ICITech), Salatiga, Indonesia, 23–25 September 2021; pp. 231–236.
47. Gkatzamanidou, M.; Papavasileiou, K.; Papavasileiou, V.; Merkouris, C.; Karras, A. 1191P Efficient lung cancer stage prediction and outcome informatics with Bayesian deep learning and MCMC method. *Ann. Oncol.* **2024**, *35*, S769. [CrossRef]
48. Amanzholova, A.; Coskun, A. Enhancing cancer stage prediction through hybrid deep neural networks: A comparative study. *Front. Big Data* **2024**, *7*, 1359703. [CrossRef] [PubMed]
49. Yu, X.; Cao, S.; Zhou, Y.; Yu, Z.; Xu, Y. Co-expression based cancer staging and application. *Sci. Rep.* **2020**, *10*, 10624. [CrossRef]

50. National Cancer Institute at the National Institutes of Health. The Cancer Genome Atlas Program: Genomic Data Commons Data Portal. Available online: <https://portal.gdc.cancer.gov/> (accessed on 15 August 2022).
51. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2012**, *2*, 401–404. [CrossRef] [PubMed]
52. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **2013**, *6*, p11. [CrossRef]
53. DeBerardinis, R.J.; Chandel, N.S. Fundamentals of cancer metabolism. *Sci. Adv.* **2016**, *2*, e1600200. [CrossRef]
54. Tan, S.Y.G.L.; van Oortmarssen, G.J.; de Koning, H.J.; Boer, R.; Habbema, J.D.F. The MISCAN-Fadia continuous tumor growth model for breast cancer. *J. Natl. Cancer Inst. Monogr.* **2006**, *36*, 56–65. [CrossRef]
55. Lee, H.; Choi, H. Investigating the Clinico-Molecular and Immunological Evolution of Lung Adenocarcinoma Using Pseudotime Analysis. *Front. Oncol.* **2022**, *12*, 828505. [CrossRef] [PubMed]
56. Bazyari, M.J.; Saadat, Z.; Firouzjaei, A.A.; Aghaee-Bakhtiari, S.H. Deciphering colorectal cancer progression features and prognostic signature by single-cell RNA sequencing pseudotime trajectory analysis. *Biochem. Biophys. Rep.* **2023**, *35*, 101491. [CrossRef]
57. Zhao, N.; Yang, K.; Yang, G.; Chen, D.; Tang, H.; Zhao, D.; Zhao, C. Aberrant expression of clock gene period and its correlations with the growth, proliferation and metastasis of buccal squamous cell carcinoma. *PLoS ONE* **2013**, *8*, e55894.
58. Abou Youssef, T.; Tanguay, S. Natural history and management of small renal masses. *Curr. Oncol.* **2009**, *16*, S2–S7. [CrossRef]
59. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef]
60. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091–D1097. [CrossRef]
61. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; et al. DrugBank 3.0: A comprehensive resource for ‘omics’ research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035–D1041. [CrossRef]
62. Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906. [CrossRef] [PubMed]
63. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672. [CrossRef] [PubMed]
64. Harding, S.D.; Armstrong, J.F.; Faccenda, E.; Southan, C.; Alexander, S.P.H.; Davenport, A.P.; Pawson, A.J.; Spedding, M.; Davies, J.A.; NC-IUPHAR. The IUPHAR/BPS guide to PHARMACOLOGY in 2022: Curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Res.* **2022**, *50*, D1282–D1294. [CrossRef]
65. Stratton, M.R.; Campbell, P.J.; Futreal, P.A. The cancer genome. *Nature* **2009**, *458*, 719–724. [CrossRef]
66. Rajendran, B.K.; Deng, C.X. Characterization of potential driver mutations involved in human breast cancer by computational approaches. *Oncotarget* **2017**, *8*, 50252–50272. [CrossRef] [PubMed]
67. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
68. Bridle, J.S. Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters. In Proceedings of the Advances in Neural Information Processing Systems 2 (NIPS 1989), Denver, CO, USA, 27–30 November 1989.
69. Hancock, J.T.; Khoshgoftaar, T.M. Survey on categorical data for neural networks. *J. Big Data* **2020**, *7*, 28. [CrossRef]
70. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
71. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef]
72. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **2013**, *4*, 627–635.
73. Corbacioglu, S.K.; Aksel, G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turk. J. Emerg. Med.* **2023**, *23*, 195–198. [CrossRef] [PubMed]
74. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendi, M.C.; Kim, J.; Reardon, B.; et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **2018**, *173*, 371–385.e18. [CrossRef]
75. Iranzo, J.; Martincorena, I.; Koonin, E.V. Cancer-mutation network and the number and specificity of driver mutations. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E6010–E6019. [CrossRef] [PubMed]
76. López-García, G.; Jerez, J.M.; Franco, L.; Veredas, F.J. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PLoS ONE* **2020**, *15*, e0230536. [CrossRef]
77. Kwon, C.; Park, S.; Ko, S.; Ahn, J. Increasing prediction accuracy of pathogenic staging by sample augmentation with a GAN. *PLoS ONE* **2021**, *16*, e0250458. [CrossRef]

78. Malhotra, K.; Fenton, J.J.; Duberstein, P.R.; Epstein, R.M.; Xing, G.; Tancredi, D.J.; Hoerger, M.; Gramling, R.; Kravitz, R.L. Prognostic accuracy of patients, caregivers, and oncologists in advanced cancer. *Cancer* **2019**, *125*, 2684–2692. [[CrossRef](#)] [[PubMed](#)]
79. Volkovova, K.; Bilanicova, D.; Bartonova, A.; Letašiová, S.; Dusinska, M. Associations between environmental factors and incidence of cutaneous melanoma. *Environ. Health* **2012**, *11*, S12.
80. Parkin, D.M.; Boyd, L.; Walker, L.C. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *Br. J. Cancer* **2011**, *105*, S77–S81. [[CrossRef](#)]
81. Spratt, D.E.; Chan, T.; Waldron, L.; Speers, C.; Feng, F.Y.; Ogunwobi, O.O.; Osborne, J.R. Racial/Ethnic Disparities in Genomic Sequencing. *JAMA Oncol.* **2016**, *2*, 1070–1074. [[CrossRef](#)] [[PubMed](#)]
82. Pharoah, P.D.; Guilford, P.; Caldas, C.; The International Gastric Cancer Linkage Consortium. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* **2001**, *121*, 1348–1353. [[CrossRef](#)]
83. Corso, G.; Veronesi, P.; Sacchini, V.; Galimberti, V. Prognosis and outcome in CDH1-mutant lobular breast cancer. *Eur. J. Cancer Prev.* **2018**, *27*, 237–238. [[CrossRef](#)]
84. Corso, G.; Intra, M.; Trentin, C.; Veronesi, P.; Galimberti, V. CDH1 germline mutations and hereditary lobular breast cancer. *Fam. Cancer* **2016**, *15*, 215–219. [[CrossRef](#)] [[PubMed](#)]
85. Chen, W.S.; Bindra, R.S.; Mo, A.; Hayman, T.; Husain, Z.; Contessa, J.N.; Gaffney, S.G.; Townsend, J.P.; Yu, J.B. CDKN2A Copy Number Loss Is an Independent Prognostic Factor in HPV-Negative Head and Neck Squamous Cell Carcinoma. *Front. Oncol.* **2018**, *8*, 95.
86. Gadhikar, M.A.; Zhang, J.; Shen, L.; Rao, X.; Wang, J.; Zhao, M.; Kalu, N.N.; Johnson, F.M.; Byers, L.A.; Heymach, J.; et al. CDKN2A/p16 Deletion in Head and Neck Cancer Cells Is Associated with CDK2 Activation, Replication Stress, and Vulnerability to CHK1 Inhibition. *Cancer Res.* **2018**, *78*, 781–797.
87. Zhou, C.; Shen, Z.; Ye, D.; Li, Q.; Deng, H.; Liu, H.; Li, J. The Association and Clinical Significance of CDKN2A Promoter Methylation in Head and Neck Squamous Cell Carcinoma: A Meta-Analysis. *Cell. Physiol. Biochem.* **2018**, *50*, 868–882. [[CrossRef](#)]
88. Martínez-Sáez, O.; Chic, N.; Pascual, T.; Adamo, B.; Vidal, M.; González-Farré, B.; Sanfeliu, E.; Schettini, F.; Conte, B.; Brasó-Maristany, F.; et al. Frequency and spectrum of PIK3CA somatic mutations in breast cancer. *Breast Cancer Res.* **2020**, *22*, 45. [[CrossRef](#)]
89. Chen, F.; Liu, J.; Song, X.; DuCote, T.J.; Byrd, A.L.; Wang, C.; Brainson, C.F. EZH2 inhibition confers PIK3CA-driven lung tumors enhanced sensitivity to PI3K inhibition. *Cancer Lett.* **2022**, *524*, 151–160. [[PubMed](#)]
90. Anderson, E.J.; Mollon, L.E.; Dean, J.L.; Warholak, T.L.; Aizer, A.; Platt, E.A.; Tang, D.H.; Davis, L.E. A Systematic Review of the Prevalence and Diagnostic Workup of PIK3CA Mutations in HR+/HER2– Metastatic Breast Cancer. *Int. J. Breast Cancer* **2020**, *2020*, 3759179. [[CrossRef](#)]
91. Schon, K.; Tischkowitz, M. Clinical implications of germline mutations in breast cancer: TP53. *Breast Cancer Res. Treat.* **2018**, *167*, 417–423. [[CrossRef](#)]
92. Zhu, G.; Pan, C.; Bei, J.-X.; Li, B.; Liang, C.; Xu, Y.; Fu, X. Mutant p53 in Cancer Progression and Targeted Therapies. *Front. Oncol.* **2020**, *10*, 595187. [[CrossRef](#)]
93. Rivlin, N.; Brosh, R.; Oren, M.; Rotter, V. Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. *Genes Cancer* **2011**, *2*, 466–474. [[CrossRef](#)] [[PubMed](#)]
94. André, F.; Ciruelos, E.; Rubovszky, G.; Campone, M.; Loibl, S.; Rugo, H.S.; Iwata, H.; Conte, P.; Mayer, I.A.; Kaufman, B.; et al. Alpelisib for PIK3CA-Mutated, Hormone Receptor-Positive Advanced Breast Cancer. *N. Engl. J. Med.* **2019**, *380*, 1929–1940.
95. Dreyling, M.; Santoro, A.; Mollica, L.; Leppä, S.; Follows, G.; Lenz, G.; Kim, W.S.; Nagler, A.; Dimou, M.; Demeter, J.; et al. Long-term safety and efficacy of the PI3K inhibitor copanlisib in patients with relapsed or refractory indolent lymphoma: 2-year follow-up of the CHRONOS-1 study. *Am. J. Hematol.* **2020**, *95*, 362–371.
96. Soria, J.C.; LoRusso, P.; Bahleda, R.; Lager, J.; Liu, L.; Jiang, J.; Martini, J.-F.; Macé, S.; Burris, H. Phase I dose-escalation study of pilaralisib (SAR245408, XL147), a pan-class I PI3K inhibitor, in combination with erlotinib in patients with solid tumors. *Oncologist* **2015**, *20*, 245–246. [[CrossRef](#)]
97. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
98. Rosenberg, R.N.; Lambracht-Washington, D.; Yu, G.; Xia, W. Genomics of Alzheimer Disease: A Review. *JAMA Neurol.* **2016**, *73*, 867–874. [[CrossRef](#)] [[PubMed](#)]
99. Sharma, N.; Cutting, G.R. The genetics and genomics of cystic fibrosis. *J. Cyst. Fibros.* **2020**, *19*, S5–S9. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.