Gap-App: A sex-distinct AI-based predictor for pancreatic ductal adenocarcinoma survival as a web application open to patients and physicians

Anuj Ojha[1,3], Shu-Jun Zhao[1,3], Basil Akpunonu[1], Jian-Ting Zhang[2], Kerri A. Simo[4,5], Jing-Yuan Liu[1,2,3,*]

[1] Department of Medicine, College of Medicine, University of Toledo, Toledo, OH, USA

[2] Department of Cell and Cancer Biology, College of Medicine, University of Toledo, Toledo, OH, USA

[3] Department of Bioengineering, College of Engineering, University of Toledo, Toledo, OH, USA

[4] Department of Surgery, College of Medicine, University of Toledo, Toledo, OH, USA

[5] ProMedica Health System, ProMedica Cancer Institute, Toledo, OH, USA

* Correspondence: jingyuan.liu@utoledo.edu; Tel.: +1-419-383-4149

Abstract

In this study, using RNA-Seq gene expression data and advanced machine learning techniques, we identified distinct gene expression profiles between male and female pancreatic ductal adenocarcinoma (PDAC) patients. Building upon this insight, we developed sex-specific 3-year survival predictive models along with a single comprehensive model. These sex-specific models outperformed the single general model despite the smaller sample sizes. We further refined our models by using the most important features extracted from these initial models. The refined sex-specific predictive models achieved improved accuracies of 92.62% for males and 91.96% for females, respectively, versus an accuracy of 87.84% from the refined comprehensive model, further highlighting the value of sex-specific analysis. Based on these findings, we created Gap-App, a web application that enables the use of individual gene expression profiles combined with sex information for personalized survival predictions. Gap-App, the first online tool aiming to bridge the gap between complex genomic data and clinical application and facilitating more precise and individualized cancer care, marks a significant advancement in personalized prognosis. The study not only underscores the importance of acknowledging sex differences in personalized prognosis, but also sets the stage for the shift from traditional one-size-fits-all to more personalized and targeted medicine. The GAP-App service is freely available at www.gap-app.org.

Keywords: pancreatic ductal adenocarcinoma (PDAC), RNA-Seq, machine learning, feature selection, gene expression, survival, personalized prognosis, Gap-App

1. Introduction

Cancer remains one of the leading causes of death worldwide, with an estimation of 2,001,140 new cases (1,029,080 in male and 972,060 in female) and 611,720 related deaths (322,800 in male and 288,920 in female) in the United States in 2024 [1]. Pancreatic cancer, accounting for about 3% of all cancer cases, disproportionately contributes to approximately 7% of all cancer-

related deaths [2]. The 5-year survival rate for pancreatic cancer was reported alarmingly low at 13% as of 2024 [3]. Specifically, there are an estimated 66,440 new cases of pancreatic cancer (34,530 in males and 31,910 in females) and 51,750 related deaths (27,270 in males and 24,480 in females) in the United States in 2024 [3], ranking the 10th and 8th most common cancer in men and women, respectively, but the 3rd leading cause of cancer-related deaths in both men and women [4]. Treatment options for pancreatic cancer can include surgery, chemotherapy, radiation therapy, immunotherapy, or a combination of these treatments. However, pancreatic cancer is a resistant disease Early detection and continued research play crucial roles in enhancing the prognosis and outcomes for individuals facing pancreatic cancer.

Personalized medicine, which customizes healthcare to the individual patient's genetic profile, offers new hope for those battling pancreatic cancer. The practice of personalized medicine involves selecting the most suitable diagnostic, therapeutic and preventive strategies for a particular patient [5] by probing specific changes in genes or gene expression levels, which are medical signs of a disease at the molecular level known as individual biomarkers [6, 7]. This is made possible by the advent of next-generation sequencing (NGS) technology that enables the rapid sequencing of multiple genes simultaneously. RNA sequencing (RNA-Seq), a prominent NGS method, has become instrumental in quantifying gene expression levels in the RNA pool, also known as transcriptome, in a biological sample. This provides insights on gene activity and regulation, leading to a deeper understanding of the cancer etiology and progression of a specific patient.

The field of cancer research has greatly benefited from the use of RNA-Seq. Studies on differential gene expression profiles for cancer biomarkers, cancer drug resistance, cancer heterogeneity and evolution have all been made possible by this technology [8]. Recent research underscores effectiveness of RNA-Seq in identifying a wide range of genetic abnormalities in diseases like acute leukemia [9, 10], offering insights into the function of differentially expressed genes and their potential molecular mechanisms [11]. Consequently, RNA-Seq has become a cornerstone in both the study and treatment of cancer, enabling the identification of an ever-increasing number of differentially expressed genes and the discovery of novel potential cancer biomarkers [12, 13]. Moreover, the decreasing cost of RNA-seq sequencing, driven by technological innovations and heightened competition, is making RNA-Seq services more affordable and accessible to patients and physicians. Commercial services offering RNA-Seq testing directly to individuals are becoming increasingly affordable, promising to democratize access to personalized medicine. Given these trends, it is foreseeable that RNA-Seq, along with other NGS techniques, will soon become commonplace in clinical settings, heralding a new era in the personalized treatment of cancer.

The advent of the NGS era in clinical settings is further propelled by rapid advancements in Artificial Intelligence (AI). The intricate and high-dimensional data produced by RNA-Seq and similar technologies necessitate sophisticated AI algorithms to uncover patterns and extract features crucial for making accurate predictions and informed decisions [14, 15]. Recent strides in AI have significantly enhanced our ability to predict medical outcomes, including

survival times, by analyzing NGS data. This has led to the development of clinical decision support systems that utilize machine learning techniques to assess survival times and recommend appropriate treatments based on these analyses [16-18].

In our study, we formulated a 3-year survival predictive model for pancreatic ductal adenocarcinoma (PDAC), the most common type of pancreatic cancer, by utilizing RNA-seq gene expression datasets from the GDC TCGA pancreatic cancer cohort (Study abbreviation: PAAD) [19]. Our decision to focus on a 3-year survival model, as opposed to a 5-year survival model, stems from the notably low average 5-year survival rate for pancreatic cancer [3, 20]. We found that despite smaller sample size, sex-dependent predictive models outperformed a single overall model in predicting patients' 3-year survival, highlighting the essential need to treat male and female patients differently. In addition, leveraging these models, we developed a web application named Gap-App, designed to be accessible to both patients and physicians. We envisage that this web application could offer valuable insights for clinical decision-making, aiding physicians in tailoring treatment plans based on the projected prognosis. Conversely, Gap-App stands to benefit patients and their families by providing a clearer outlook on the expected disease trajectory, allowing for better financial and emotional preparation for future healthcare needs. Understanding realistic survival expectations helps patients and their families to mentally brace for forthcoming challenges and to focus on quality of life [21].

2. Methods

2.1. Data Acquisition and Preprocessing: The gene expression profiles and clinical information for all 178 pancreatic ductal adenocarcinoma (PDAC) patients were downloaded from the GDC TCGA database via the UCSC Xena browser [22]. Censoring was applied to these patients, and all censored patients whose 3-year survival outcome was unknown were excluded when calculating survival rates and building survival predictive models.

In addition to the demographic and surgical information, gene expression profiles of the cohort were obtained in the $\log_2$(FPKM+1) format from the GDC TCGA database. Furthermore, in order to perform Differential Gene Expression (DGE) analysis between cancer and normal tissues, it was essential to obtain the gene expression data from normal pancreas tissue in addition to the primary tumor samples. Along with data from 4 patients within the GDC TCGA cohort, gene expression profiles from an additional 172 normal samples were accessed using the Genotype-Tissue Expression (GTEx) cohort dataset through Toil Hub [22, 23]. Gene expression profiles in $\log_2$(counts+1) format, which is the appropriate format for DGE analysis, were downloaded from both databases. Subsequently, the data were meticulously cleaned, involving the removal of all duplicates and genes that have missing expression values in more than 75% of the patients.

2.2. Model building to predict sex of a patient sample: Machine learning (ML) models were developed by employing random forest, logistic regression, and linear support vector machine algorithms to predict the sex of a sample based on RNA-Seq data of 19,318 or 19,253 protein-coding genes with or without Y-chromosome genes, respectively. Model performance was evaluated using accuracy, precision, recall, and F1 score with 5-fold cross-validation.

2.3. Feature selection and gene filtration for initial survival model building: In the original dataset for each patient, transcript reads for approximately 60,000 genes are available, representing the total features accessible for feeding our ML models. However, an excess of features can lead to data overfitting that ends up with a model that captures random fluctuations and noises rather than the authentic patterns of a given training set, leading to suboptimal predictions on new dataset. The risk of overfitting escalates particularly when the ML models are fed with too many features from datasets with small sample sizes, which is the predicament we face. To mitigate this risk, we conducted initial gene filtration based on three criteria (Figure 1):

1. The gene's expression level must significantly impact survival ($p<0.05$), as determined by hazard ratio (HR) using the Cox proportional-hazards (Cox PH) model [24, 25].
2. The gene's fold change must be greater than 2 and significant (FDR<0.05), as determined by using edgeR [26] through differential gene expression (DGE) analysis between normal and primary tumor tissue samples.
3. A gene must produce a known protein product and be present in the largest cluster of the gene group as determined by STRING database V.12, which was visualized by Cytoscape version 3.10.0 [27].

Only genes meeting all three criteria were retained for further analysis.

In detail, we first evaluated the impact of each gene on survival by employing Cox PH model. Patients were divided into high and low expression groups for each gene, based on a cutoff expression value that was determined by the maximum selected rank statistics using the survminer and maxstat package in R [28]. Then hazard ratio of high versus low expression of each gene was calculated by Cox PH model using the 'survival' package in R (R version 4.3.0) [24]. Genes with a p-value <0.05 were considered impactful and retained, resulting in 18,040, 13,965 and 11,181 significant genes for overall, males, and female groups, respectively (Figure 1).

Concurrently, all genes were filtered by DGE analysis using edgeR, a Bioconductor software in R, to compare the different expression level of genes between the tumor and normal tissue groups. Genes that have a logarithmic fold change ($\log_2(FC)$) >1 or <-1 while simultaneously having False Discovery Rate (FDR) value <0.05 were retained, yielding 17,715, 17,828 and 17,554 significant genes for overall, male, and female groups, respectively, for further downstream analysis (Figure 2).

Finally, genes that simultaneously passed the above two criteria were then analyzed and clustered using STRING database V.12 and visualized by Cytoscape 3.10.0. Only genes present in the largest cluster were retained, leading to a list of 5,098, 3,839 and 3,219 genes for our initial ML model building for overall, male and female groups, respectively.

Additionally, another set of models for overall, male, and female groups were built by adding patients' age, tumor stage and type of surgery to the gene expression features from the above gene lists.

2.4. Initial Model Training and Validation: Two types of ML models, a single comprehensive model for the entire patient cohort and sex-dependent models with distinct models for males and females, were built. Five ML algorithms: Random Forest (RF), AdaBoost, Logistic Regression (LR), Decision Tree (DT), and Linear Support Vector Machine (SVM), were

employed using the scikit-learn package in Python [29]. Patient data were split into training and testing sets at a ratio of 4:1 and performed repeated stratified 5-fold cross-validation. Parameter tuning and optimization were performed across all models. The best models were selected based on mean accuracy, precision, recall, and F1 score from repeated stratified 5-fold cross-validation.

2.5. Feature Extraction and Model Refinement: To further minimize the risk of overfitting, a second round of feature selection was performed using the '*feature_importances_*' attribute from the best initial ML models in the sickit-learn package. Features with a score of zero, indicating no relevance, were eliminated, resulting in refined gene lists of 999, 387, and 108 genes for the overall, male, and female groups, respectively. These refined features were used to train the five ML algorithms again in the same way as listed above and evaluated using the same strategy.

2.6. Web Application Development: Leveraging sex-specific ML models, a web application named Gap-App was developed. The trained ML models were saved as pickle files (.pkl) and integrated into a Flask-based web application. These trained models in pickle files are available in the GitHub repository. User requests are handled by a Web Server Gateway Interface (WSGI) server, which serves the Flask application. The primary interface is designed to be intuitive and user-friendly, allowing users to select their sex, upload RNA-Seq gene expression data, and receive personalized survival predictions based on the appropriate ML model.

3. Results

3.1. Data origin and description

The gene expression profiles and clinical information for all 178 PDAC patients with an initial pathologic diagnosis were downloaded from the GDC TCGA database. Table 1 presents the demographic distribution including sex, age, race, and tumor stage at diagnosis within the cohort. The cohort consisted of 80 (44.94%) females and 98 (55.06%) males, and 155 (87.10%) Caucasians, 11 (6.18%) Asians and 6 (3.37%) African Americans. The mean age is 65 years for both male and female subgroups and the entire cohort. Nearly all patients underwent surgical procedures, with only 4 cases not reported (Table 1). The Whipple procedure was the most performed, with 135 patients (75.84%, 9 survivors, 74 non-survivors, 52 censored), followed by Distal Pancreatectomy with 23 patients (12.92%, 5 survivors, 11 non-survivors, 7 censored). Total Pancreatectomy was less common, involving only 3 males (1.69% of total patients, 0 survivors, 1 non-survivor, 2 censored). Finally, other methods were used in 13 patients (7.30%, 5 survivors, 1 non-survivor, 7 censored). Given the nearly unequal distribution of survivors and non-survivors with the most common Whipple procedure (higher number of non-survivors compared to very few numbers of survivors), along with varied survival outcomes from very small samples with less frequent surgeries, determining whether good survival is directly linked to a specific type of surgery is very challenging. This complexity indicates that factors beyond the choice of surgical method may significantly influence survival rates.

Additionally, Table 1 also reports the pathological diagnosis of tumor stages ranging from stage I to IV, with stage IIb (65.73%) predominating the cohort. The 3-year survival outcome for different tumor stages of each sex in this cohort were calculated and are displayed in Table 2.

In the male subgroup, the lowest survival rate was observed in patients with Stage IIb (10.81%) followed by stage Ib (25.00%). Patients with stage IIa exhibited a higher 3-year survival rate of 30.00%. For females, the lowest 3-year survival rate was noted in patients with stage IIb (12.50%), followed by stage IIa (33.33%), and then stage Ib (60.00%) in ascending order. It was observed that, in both sex groups, patients with stage IIb had lower survival rate compared to most other stages. Additionally, stage IIb patients demonstrated the widest range in their survival time (male: 12 – 1059 days vs. female: 31 – 913 days). These findings underscore the critical need for developing a more effective predictive survival tool for pancreatic cancer patients than the current clinical practice of relying on the traditional pathological criteria, and patient characteristics, which suffered from methodological limitations or fell short in predictive accuracy as reviewed by Strijker et al. in 2019 [30].

To perform the differential gene expression analysis, along with the available data of normal pancreas tissues from 4 patients (comprising 2 males and 2 females) within the GDC TCGA cohort, we utilized the Genotype-Tissue Expression (GTEx) cohort dataset using Toil Hub [22, 23]. This enabled us to access gene expression profile from an additional 172 normal samples (97 males and 75 females).

3.2. Primary tumor samples from males and females have different gene expression patterns that can be predicted by ML models

Prior to initiating the development of models for survival prediction, we posited a question: Do males and females exhibit divergent gene expression patterns in their primary tumors, potentially influencing survival outcomes differently, thereby necessitating the construction of sex-specific predictive models? In pursuit of an answer, we set out to develop a machine learning (ML) model capable of distinguishing the sex of a sample based on the RNA-Seq data of 19,318 protein-coding genes as cataloged by the STRING database V.12 [31]. Initially, we employed three AI models, namely random forest, logistic regression, and linear support vector machine, to predict the sample's sex from gene expression profile. Model performance was assessed by accuracy, precision, recall, and F1 score. The averaged values of these metrics from 5-fold cross-validation are listed in Table 3. Remarkably, all three ML models achieved high accuracy exceeding 95%. Subsequent examination of the most powerful predictive genes revealed that this high accuracy likely stems from the inclusion of Y-chromosome-specific genes. To determine whether males and females demonstrate different expression patterns for their shared genes within their primary tumors, we refined our predictive models to exclude genes located on the Y-chromosome, focusing on 19,253 protein-coding genes. Among all 3 machine learning models evaluated, the random forest algorithm emerged as the most effective, achieving an impressive accuracy of 90.32% (Table 3) and an area under the Receiver Operating Characteristic (ROC) curve (AUC) of 0.95 (Figure 1B). These findings strongly indicate that male and female patients exhibit unique gene expression patterns in their PDAC primary tumor, which can be captured by machine learning techniques. The observed differences in expression patterns underscore the potential benefit of developing sex-specific survival models to enhance the accuracy of patient survival predictions. To explore this

hypothesis further and to strive for greater model accuracy, we constructed a comprehensive ML model alongside two separate models for males and females, respectively.

3.3. Initial ML building suggests sex-dependent ML models outperform one unified ML approach

To determine whether accounting for sex differences enhances the accuracy of 3-year survival prediction of PDAC patients, we embarked on building two types of ML models: a single comprehensive ML model predicting the survival of the entire patient cohort and sex-dependent ML models, which include distinct models for males and females, respectively. The patient data was split into training and testing sets at a ratio of 4:1, and repeated stratified 5-fold cross-validation was performed. Gene expression values in the $\log_2$(FPKM+1) format from the final gene lists consisting of 5,098, 3,839 and 3,219 genes for overall, female, and male groups, respectively, were utilized to feed five ML algorithms that are Random Forest (RF), AdaBoost, Logistic Regression (LR), Decision Tree (DT), and Linear Support Vector Machine (SVM). Following parameter tuning and optimization, these initial models were assessed by their accuracy, precision, recall, and F1 score, with the mean and range values from repeated stratified 5-fold cross-validation presented in Table 4. The best models for each group are in bold as indicated by these mean metrics. For the single overall model, Random Forest exhibited the highest mean accuracy (87.84%), followed by AdaBoost and Logistic Regression (86.16% for both). In the sex-dependent ML models, the highest accuracy for the male group is AdaBoost (88.47%) followed by Random Forest (87.74%); and for the female group, the best is also AdaBoost (88.94%), followed by Random Forest (88.62%). Therefore, despite a reduction in sample sizes by nearly half for both sexes, the sex-dependent ML models demonstrated superior mean accuracy with AdaBoost, achieving 88.47% and 88.94% for male and female, respectively, surpassing the highest accuracy, 87.84%, of the singular overall approach.

In agreement with the mean accuracy values, the AUC values of the ROC curves indicated that Random Forest, with an AUC value of 0.73, led overall (Figure 4A), while AdaBoost, with AUC values of 0.76 for males and 0.85 for females, excelled sex-dependent ML models (Figure 4B-C). For convenient comparison, ROC curves of the best single overall model and sex-dependent models are plotted in figure 4D, clearly showing that sex-specific approach is superior to the single model approach.

3.4. Added patients' clinical characteristic features failed to improve ML models

In an effort to further improve the ML models, we built another set of models for overall, male, and female groups by adding patients' age, tumor stage and type of surgery to the existing gene expression features. The model performance metrics are presented in Table 5. For the overall group, the best single model was achieved using Random Forest, with a mean accuracy of 87.84% and an AUC value of 0.68 (Figure 5A). For the sex-dependent ML models, AdaBoost performed the best for males, with an averaged accuracy of 87.74% and an AUC value of 0.73 (Figure 5B). Similarly, for females, AdaBoost achieved the highest mean accuracy of 88.94% and an AUC value of 0.84 (Figure 5C). For easy comparison, the ROC

curves of the best overall model and sex-dependent models are plotted in Figure 5D. Clearly, although the trend of sex-specific models outperforming the single overall approach remains, the addition of patients' characteristic features did not surpass the performance of models using only RNA-Seq gene reads. Consequently, the initial models developed using only the gene expression features were used for further feature analysis and refinement of the models by feature extraction for WebApp development.

3.5. Models are further enhanced by selecting optimal features from the best preliminary ML models

To further minimize the risk of overfitting in our models, we explored the possibility of reducing the number of features. This approach not only boosts model performance but also decreases computational time and memory requirements [32], simplifies understanding and interpretation of the model, and leads to deeper insights into the intrinsic relationships within the data [33]. This selection process may also enhance model robustness and stability by eliminating irrelevant features and focusing on those that are most impactful, thus reducing sensitivity to minor data variations [34].

We employed the '*feature_importances_*' attribute from the best initial ML models to score the importance of each feature and only features with a non-zero score were kept. This filtering process resulted in the selection of 999, 387 and 108 genes for constructing subsequent ML models for overall, male, and female groups, respectively. Remarkably, using the gene expressions from these refined gene lists led to observable performance improvements in the new set of ML models. The results from repeated stratified 5-fold cross-validation of these refined models are detailed in Table 6. The best models for overall, male, and female are highlighted in bold and marked by an asterisk (*), according to these metrics.

For the overall model, Random Forest achieved the highest mean accuracy of 87.84% with an AUC of 0.75 (Figure 6A). In the sex-dependent categories, LinearSVC outperformed others for males with an average accuracy of 92.62% and an AUC of 0.95 (Figure 6B), while Logistic Regression achieved the highest mean accuracy of 91.96% and an AUC of 0.96 for females (Figure 6C). For easy comparison, ROC curves of the top-performing overall model and the sex-dependent models are depicted in figure 6D, illustrating the superiority of the sex-specific approach once again. Additionally, compared with the initial models, the refined ML models showed tremendous improvement as indicated by not only the mean performance metrics but also the narrower ranges of these metrics, suggesting a more robust and stable predictive power. As such, the best sex-dependent models with accuracies of 92.62% and 91.96% for male and female groups, respectively, were selected for further WebApp development. These selected models are labeled by a dagger mark (†) in Table 6.

3.6. The developed Gap-App web application

Leveraging these sex-specific ML models, we developed a web application named Gap-App, an acronym for "Gender-distinct AI-based Predictor as a web application", where gender refers to gender at birth. When a user visits the main webpage, their request is handled by a Web Server Gateway Interface (WSGI) server, which is responsible for serving the Flask-based

application [35]. The webpage itself is constructed using HTML, CSS, and JavaScript to enable dynamic interaction between the frontend and the Flask backend, ensuring content is delivered efficiently and accurately to the user's web browser (Figure 7A) [36]. The primary interface is designed to be intuitive and user-friendly (Figure 7B). Upon starting, users select their sex via a radio button which is a critical step, as the prediction model tailors its analysis accordingly. Next, they are asked to upload their RNA-Seq gene expression data in FPKM format, which will be converted to $\log_2$(FPKM+1) transformed form automatically in our application for the ensuing analysis. It is important to note that the gene expression data should be from an RNA-Seq test targeting the whole transcriptome rather than certain isolated genes. Users must then accept the Terms and Agreements to proceed, affirming their usage of the service is bound by the terms and conditions of the Agreement. After submission, Gap-App first verifies the presence of all genes. If no genes are missing, the application, utilizing Python and Flask, selects the appropriate machine learning (ML) model based on sex to predict the 3-year survival outcome, which is displayed in a result page, offering personalized insights into their prognosis (Figure 7C).

Gap-App can also address missing genes by assigning a gene expression value of 0, which may more accurately reflect the reality than using mean or median as our predictions are based on whole transcriptome RNA-Seq. If a gene is absent, assuming it undetectable is reasonable. However, if more than 10% of the genes are missing, Gap-App will not make a prediction but will instead issue an error message, prompting users to verify if their dataset is from a whole transcriptome RNA-Seq. GAP-App is freely available at www.gap-app.org.

Conclusion and 1234**Discussion**

Like many other diseases, PDAC exhibits comparable disease incidence and survival rates in males and females. However, this statistical similarity may not imply identical molecular disease mechanisms in both sexes. Our supervised machine learning model can discern different male and female gene expression patterns with high accuracy, indicating distinct gene expression patterns between these two sexual groups. Most strikingly, despite a reduction in sample sizes by nearly half for both sexes, sex-specific ML models significantly outperformed the one-size-fits-all single ML approach. These findings underscore the importance of acknowledging sex-dependent differences in gene expression patterns, which may point to divergent molecular pathways and disease progression in male and female PDAC patients. Building upon these research findings, we developed a novel, user-friendly web application to predict 3-year survival for PDAC patients.

Previously, many survival prediction models for pancreatic cancer have been developed by incorporating pathologic criteria such as tumor stage, grade, and other patient characteristics [30]. However, these traditional survival prediction models often vary widely and offer limited accuracy, with some, like the MSKCCPAN model, severely underestimating survival rates. Interestingly, our attempts to integrate these traditional predictors as additional features into our models revealed a fascinating paradox: rather than improving, they slightly compromised the model’s performance as can be seen in Table 4 and 5, Figure 4 and 5, echoing the limited prediction accuracy based on these clinical parameters alone. This unexpected outcome

necessitates a critical reevaluation of conventional prognostic frameworks in the context of molecular and genetic data. It suggests that while clinical parameters such as age, tumor stage and type of surgery performed are undeniably influential, they might not capture the full spectrum of biological variance at the genetic level. Moreover, the observed discrepancy could be interpreted as evidence of the intricate interplay between genetic factors and traditional clinical predictors, pointing towards a multifaceted disease ontology where genetic predispositions and expressions provide additional, nuanced insights into patient prognosis beyond what can be gleaned from clinical observations alone. This revelation emphasizes the importance of developing more sophisticated, multimodal predictive models that harness the potential of both genetic and clinical data to provide a more accurate, personalized prognosis. Finally, in light of these findings, future research should aim to delve deeper into the mechanisms through which sex dictates different molecular pathways and disease progression in male and female patients and subsequent determination of sex-specific biomarkers in PDAC. It is noteworthy that personalized treatment for PDAC has been ineffective [37] possibly due to lack of considerations of sex and other factors. Such endeavors could pave the way for truly personalized medicine, where treatment strategies are tailored not only based on observable clinical factors but are also tuned to reflect the underlying genetic landscape of each patient.

While our study represents a significant advancement in the predictive modeling of pancreatic cancer survival, it is important to acknowledge the inherent challenges impacting model accuracy, in addition to limitations related to algorithms, as reviewed by Sharma et al. in AI-based predictions of intestinal polyps and colorectal cancer [38]. Firstly, the rarity of PDAC limits the sample size for research purposes. This limitation poses significant challenges for model training and validation, as machine learning algorithms typically require large datasets to accurately learn and generalize across diverse patient populations. The small sample size risks overfitting, where models might perform well on training data but poorly on unseen data, compromising the robustness and reliability of our predictive outcomes. Secondly, the evolving landscape of cancer therapies, particularly the rapid advancement in treatments such as immunotherapy, poses another challenge. These new therapies and their combinations are continuously changing the survival outcomes for PDAC and other cancer patients [39] , introducing variability that our current model might not fully capture. As these treatments become more prevalent, the historical data used to train our models may not accurately reflect the current or future realities of PDAC treatment outcomes. This temporal shift necessitates continuous updates to the model, incorporating new data that reflect these advancements in therapy to maintain or improve predictive accuracy.

In conclusion, our findings align with the growing body of research advocating for personalized medicine in cancer care. The development of the GAP-App web application as a user-friendly tool freely available for both patients and physicians emphasizes the practical application of our research findings. By providing access to sex-specific survival predictions, this app bridges the gap between complex genetic data and clinical decision-making to foster personalized treatment strategies based on individual gene expression profiles and genetic makeup. On the

other hand, while our sex-specific predictive models mark a significant advancement in the personalized prognosis of PDAC patients, addressing the challenges outlined above is essential for further refinement and application of these models in clinical settings. Future research should focus on overcoming these obstacles, with the ultimate goal of enhancing the accuracy and applicability of predictive models in the ever-evolving landscape of cancer care.

Acknowledgements: This work was supported by the National Institutes of Health grant R01GM127656. The manuscript was proofread and grammatically refined with the assistance of ChatGPT 4.0.

CRediT author statement

Anuj Ojha: Data curation, Formal analysis, Investigation, Methodology, Validation, Original draft, review and editing
Shujun Zhao: Methodology, Software, Original draft, review and editing
Basil Akpunonu: Resources, Validation, Draft review and editing
Jian-Ting Zhang: Conceptualization, Methodology, Original draft, review and editing
Kerri A. Simo: Investigation, Methodology, Validation, Original draft, review and editing
Jing-Yuan Liu: Conceptualization, Investigation, Methodology, Funding acquisition, Resources, Supervision, Original draft - drafting, reviewing, and editing

Conflict of interest: The authors have declared no competing interest.

Notes: The trained ML models are saved as pickle files (.pkl) and can be found in GitHub repository at https://github.com/LiuLabUTMC/Gap-App. This repository is currently private but can be accessed upon request.

References

1. Siegel, R.L., A.N. Giaquinto, and A. Jemal, *Cancer statistics, 2024.* CA: A Cancer Journal for Clinicians, 2024. **74**(1): p. 12-49.
2. Society, A.C. *Key Statistics for Pancreatic Cancer*. Available from: https://www.cancer.org/cancer/types/pancreatic-cancer/about/key-statistics.html.
3. Society, A.C., *Cancer Facts & Figures 2024. Atlanta: American Cancer Society; 2024*
4. Li, O., et al., *Biological characteristics of pancreatic ductal adenocarcinoma: Initiation to malignancy, intracellular to extracellular.* Cancer Lett, 2023. **574**: p. 216391.
5. Shlyakhto, E.V., *Scientific Basics of Personalized Medicine: Realities and Opportunities.* Herald of the Russian Academy of Sciences, 2022. **92**(6): p. 671-682.
6. Raman, P., et al., *A comparison of survival analysis methods for cancer gene expression RNA-Sequencing data.* Cancer Genet, 2019. **235-236**: p. 1-12.
7. Morash, M., et al., *The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology.* J Pers Med, 2018. **8**(3).
8. Hong, M., et al., *RNA sequencing: new technologies and applications in cancer research.* Journal of Hematology & Oncology, 2020. **13**(1): p. 166.

9. Arindrarto, W., et al., *Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing.* Leukemia, 2021. **35**(1): p. 47-61.
10. Brown, L.M., et al., *The application of RNA sequencing for the diagnosis and genomic classification of pediatric acute lymphoblastic leukemia.* Blood Advances, 2020. **4**(5): p. 930-942.
11. Oshlack, A., M.D. Robinson, and M.D. Young, *From RNA-seq reads to differential expression results.* Genome biology, 2010. **11**: p. 1-10.
12. Yamada, A., et al., *A RNA-Sequencing approach for the identification of novel long non-coding RNA biomarkers in colorectal cancer.* Scientific reports, 2018. **8**(1): p. 575.
13. Zhu, L., et al., *Exosomal tRNA-derived small RNA as a promising biomarker for cancer diagnosis.* Molecular cancer, 2019. **18**(1): p. 1-5.
14. Danaee, P., R. Ghaeini, and D.A. Hendrix. *A deep learning approach for cancer detection and relevant gene identification*. in *Pacific symposium on biocomputing 2017*. 2017. World Scientific.
15. Rukhsar, L., et al., *Analyzing RNA-Seq Gene Expression Data Using Deep Learning Approaches for Cancer Classification.* Applied Sciences, 2022. **12**(4): p. 1850.
16. Bashiri, A., et al., *Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review.* Iran J Public Health, 2017. **46**(2): p. 165-172.
17. Pennathur, A., et al., *Gene expression profiles in esophageal adenocarcinoma predict survival after resection.* J Thorac Cardiovasc Surg, 2013. **145**(2): p. 505-12; discussion 512-3.
18. Ishibashi, Y., et al., *Profiling gene expression ratios of paired cancerous and normal tissue predicts relapse of esophageal squamous cell carcinoma.* Cancer Res, 2003. **63**(16): p. 5159-64.
19. Grossman, R.L., et al., *Toward a shared vision for cancer genomic data.* New England Journal of Medicine, 2016. **375**(12): p. 1109-1112.
20. Qin, L., et al., *Reversible promoter demethylation of PDGFD confers gemcitabine resistance through STAT3 activation and RRM1 upregulation.* Cancer Lett, 2023. **567**: p. 216266.
21. Glare, P.A. and C.T. Sinclair, *Palliative medicine review: prognostication.* Journal of palliative medicine, 2008. **11**(1): p. 84-103.
22. Goldman, M.J., et al., *Visualizing and interpreting cancer genomics data via the Xena platform.* Nature Biotechnology, 2020. **38**(6): p. 675-678.
23. Vivian, J., et al., *Toil enables reproducible, open source, big biomedical data analyses.* Nature Biotechnology, 2017. **35**(4): p. 314-316.
24. Therneau, T.M., *A Package for Survival Analysis in R.* 2024.
25. Therneau, T.M. and P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. 2000, New York: Springer.
26. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
27. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
28. *Kassambara, A. and Kosinski, M. (2021). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.9. https://CRAN.R-project.org/package=survminer.*
29. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
30. Strijker, M., et al., *Systematic review of clinical prediction models for survival after surgery for resectable pancreatic cancer.* Br J Surg, 2019. **106**(4): p. 342-354.
31. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-d613.
32. Li, J., et al., *Feature Selection: A Data Perspective.* ACM Comput. Surv., 2017. **50**(6): p. Article 94.

33. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**(Mar): p. 1157-1182.
34. Chen, R.-C., et al., *Selecting critical features for data classification based on machine learning methods.* Journal of Big Data, 2020. **7**(1): p. 52.
35. Olson, R.H., N.C. Kalafut, and D. Wang, *MANGEM: a web app for Multimodal Analysis of Neuronal Gene expression, Electrophysiology and Morphology.* bioRxiv, 2023.
36. Maussion, G., et al., *Auto-qPCR; a python-based web app for automated and reproducible analysis of qPCR data.* Scientific Reports, 2021. **11**(1): p. 21293.
37. Wei, H. and H. Ren, *Precision treatment of pancreatic ductal adenocarcinoma.* Cancer Lett, 2024. **585**: p. 216636.
38. Sharma, A., et al., *Artificial intelligence in intestinal polyp and colorectal cancer prediction.* Cancer Lett, 2023. **565**: p. 216238.
39. Shi, F., et al., *Improvement strategy for immune checkpoint blockade: A focus on the combination with immunogenic cell death inducers.* Cancer Lett, 2023. **562**: p. 216167.

Table 1. The demographic distribution and surgical information of the PDAC cohort

Pancreatic Cancer Patients		Male	Female	Total
Sex		98 (55.06%)	80 (44.94%)	178(100.00%)
Mean Age (range in parathesis)		65 (35 – 86) years	65 (39 – 89) years	65 (35 – 89) years
Race	Caucasian	87 (88.78%)	68 (85.00%)	155 (87.10%)
	Asian	5 (5.10%)	6 (7.50%)	11 (6.18%)
	African American	3 (3.06%)	3 (3.75%)	6 (3.37%)
	Not reported	3 (3.06%)	3 (3.75%)	6 (3.37%)
Ethnicity	Hispanic or Latino	3 (3.06%)	2 (2.50%)	5 (2.81%)
	Not Hispanic or Latino	65 (66.33%)	64 (80.00%)	129 (72.47%)
	Not reported	30 (30.61%)	14 (17.50%)	44 (24.72%)
Tumor Stage at Diagnosis	Stage I	-	1 (1.25%)	1 (0.56%)
	Stage Ia	3 (3.06%)	2 (2.50%)	5 (2.81%)
	Stage Ib	7 (7.14%)	8 (10.00%)	15 (8.43%)
	Stage IIa	18 (18.37%)	10 (12.50%)	28 (15.73%)
	Stage IIb	64 (65.31%)	53 (66.25%)	117 (65.73%)
	Stage III	1 (1.02%)	2 (2.50%)	3 (1.69%)
	Stage IV	1 (1.02%)	3 (3.75%)	4 (2.25%)
	Not reported	4 (4.08%)	1 (1.25%)	5 (2.81%)
Surgery types	Whipple	72 (73.47%) (5/36/31)*	63 (78.75%) (4/38/21)*	135 (75.84%) (9/74/52)*
	Distal Pancreatectomy	15 (15.31%) (3/7/5)*	8 (10.00%) (2/4/2)*	23 (12.92%) (5/11/7)*
	Total Pancreatectomy	3 (3.06%) (0/1/2)*	-	3 (1.69%) (0/1/2)*
	Other Methods	6 (6.12%) (2/0/4)*	7 (8.75%) (3/1/3)*	13 (7.30%) (5/1/7)*
	Not reported	2 (2.04%) (0/0/2)*	2 (2.50%) (1/0/1)*	4 (2.25%) (1/0/3)*

*The numbers in parentheses indicate the number of patients in different survival outcome in the format (survived/died/censored).

Table 2. Three-year survival outcomes of male and female subgroups by tumor stages

Stage	Total patients	Censored patients	Survivors	Deaths	survival rate (in %)	Survival range (in days)	Average survival (in days)
Male							
Stage ia	3	2	-	1	-	244	244
Stage ib	7	3	1	3	25.00%	236 - 598	444
Stage iia	18	8	3	7	30.00%	117 - 634	324
Stage iib	64	27	4	33	10.81%	12 - 1059	360
Stage iii	1	1	-	-	-	-	-
Stage iv	1	1	-	-	-	-	-
Not reported	4	2	2	-	100.00%	-	-
Total	98	44	10	44	18.52%	12 - 1059	355
Female							
Stage i	1	-	1	-	100.00%	-	-
Stage ia	2	1	-	1	-	151	151
Stage ib	8	3	3	2	60.00%	144 – 219	182
Stage iia	10	7	1	2	33.33%	299 – 381	340
Stage iib	53	13	5	35	12.50%	31 – 913	435
Stage iii	2	1	-	1	-	394	394
Stage iv	3	1	-	2	-	394 – 545	470
Not reported	1	1	-	-	-	-	-
Total	80	27	10	43	18.87%	31 - 913	413

Table 3: Performance metrics of different sex predictive ML models

Model	Accuracy (in%)	Precision	Recall	F1
Including Y-chromosome genes				
Random Forest	98.87	0.99	0.99	0.99
Logistic Regression	99.43	0.99	1.0	0.99
Linear Support Vector Machine	96.59	0.96	0.98	0.97
Excluding Y-chromosome genes				
Random Forest	90.32	0.89	0.94	0.91
Logistic Regression	80.71	0.81	0.84	0.82
Linear Support Vector Machine	63.67	0.67	0.67	0.66

Table 4. Mean and range (in parathesis) of performance metrics for various initial ML survival models using RNA-Seq reads as solely features

Predictive Models	ML model	Accuracy (in %)	Precision	Recall	F1
Single model	**Random Forest***	**87.84 (80.95 – 95.45)**	**0.87 (0.81 – 0.95)**	**1.0**	**0.93 (0.89 – 0.97)**
	AdaBoost	86.16 (68.18 – 95.45)	0.87 (0.79 – 0.95)	0.97 (0.83 – 1.0)	0.92 (0.81 – 0.97)
	Decision Tree	84.28 (63.64 – 95.45)	0.88 (0.80 – 1.0)	0.93 (0.71 – 1.0)	0.90 (0.76 – 0.97)
	Logistic Regression	86.16 (76.19 – 95.45)	0.87 (0.80 – 0.95)	0.98 (0.88 – 1.0)	0.92 (0.86 – 0.97)
	Linear Support Vector Machine	84.86 (71.43 – 95.24)	0.88 (0.79 – 0.94)	0.95 (0.83 – 1.0)	0.91 (0.83 – 0.97)
Male model	Random Forest	87.74 (80.00 – 90.91)	0.87 (0.80 – 0.90)	1.0	0.93 (0.89 – 0.95)
	AdaBoost*	**88.47 (72.73 – 90.91)**	**0.88 (0.80 – 0.90)**	**0.99 (0.89 – 1.0)**	**0.93 (0.84 – 0.95)**
	Decision Tree	76.94 (40.00 – 100.0)	0.88 (0.67 – 1.0)	0.83 (0.44 – 1.0)	0.85 (0.57 – 1.0)
	Logistic Regression	84.00 (72.73 – 90.91)	0.87 (0.80 – 0.90)	0.94 (0.78 – 1.0)	0.90 (0.82 – 0.95)
	Linear Support Vector Machine	82.87 (70.00 – 90.91)	0.87 (0.78 – 0.90)	0.93 (0.78 – 1.0)	0.90 (0.82 – 0.95)
Female model	Random Forest	88.62 (80.00 – 100.0)	0.88 (0.80 – 1.0)	1.0	0.93 (0.89 – 1.0)
	AdaBoost*	**88.94 (70.00 – 100.0)**	**0.90 (0.78 – 1.0)**	**0.98 (0.78 – 1.0)**	**0.93 (0.82 – 1.0)**
	Decision Tree	82.98 (60.00 – 90.91)	0.88 (0.80 – 1.0)	0.91 (0.62 – 1.0)	0.89 (0.71 – 0.95)
	Logistic Regression	85.64 (70.00 – 100.0)	0.89 (0.78 – 1.0)	0.94 (0.78 – 1.0)	0.91 (0.82 – 1.0)
	Linear Support Vector Machine	85.64 (70.00 – 100.0)	0.89 (0.78 – 1.0)	0.94 (0.78 – 1.0)	0.91 (0.82 – 1.0)

*Indicate the best initial survival predictive ML models developed.

Table 5. Mean and range (in parathesis) of performance metrics for various initial ML survival models using RNA-Seq reads plus patients' characteristics as features

Predictive Models	ML model	Accuracy (in %)	Precision	Recall	F1
Single model	**Random Forest***	**87.84 (80.95 – 95.45)**	**0.87 (0.81 – 0.95)**	**1.0**	**0.93 (0.89 – 0.97)**
	AdaBoost	85.96 (68.18 – 95.45)	0.87 (0.79 – 0.95)	0.97 (0.83 – 1.0)	0.92 (0.81 – 0.97)
	Decision Tree	83.77 (59.09 – 95.24)	0.86 (0.76 – 0.94)	0.96 (0.72 – 1.0)	0.90 (0.74 - 0.97)
	Logistic Regression	86.70 (76.19 – 95.45)	0.87 (0.80 – 0.95)	0.98 (0.88 – 1.0)	0.92 (0.86 – 0.97)
	Linear Support Vector Machine	84.47 (71.43 – 95.45)	0.87 (0.79 – 0.95)	0.94 (0.82 – 1.0)	0.91 (0.82 – 0.97)
Male model	Random Forest	87.74 (80.00 – 90.91)	0.87 (0.80 – 0.90)	1.0	0.93 (0.89 – 0.95)
	AdaBoost*	**87.74 (80.00 – 90.91)**	**0.88 (0.80 – 0.90)**	**0.99 (0.89 – 1.0)**	**0.93 (0.89 – 0.95)**
	Decision Tree	80.29 (60.00 – 100.0)	0.88 (0.75 – 1.0)	0.88 (0.67 – 1.0)	0.88 (0.75 – 1.0)
	Logistic Regression	84.76 (72.73 – 90.91)	0.87 (0.80 – 0.90)	0.95 (0.78 – 1.0)	0.91 (0.82 – 0.95)
	Linear Support Vector Machine	86.65 (80.00 – 90.91)	0.89 (0.80 – 1.0)	0.96 (0.87 – 1.0)	0.92 (0.89 – 0.95)
Female model	Random Forest	88.62 (80.00 – 100.0)	0.88 (0.80 – 1.0)	1.0	0.93 (0.89 – 1.0)
	AdaBoost*	**88.94 (70.00 – 100.0)**	**0.90 (0.78 – 1.0)**	**0.98 (0.78 – 1.0)**	**0.93 (0.82 – 1.0)**
	Decision Tree	81.64 (50.00 – 100.0)	0.90 (0.71 – 1.0)	0.87 (0.62 – 1.0)	0.88 (0.67 – 1.0)
	Logistic Regression	85.27 (70.00 – 100.0)	0.88 (0.78 – 1.0)	0.95 (0.78 – 1.0)	0.91 (0.82 – 1.0)
	Linear Support Vector Machine	84.91 (70.00 – 100.0)	0.88 (0.78 – 1.0)	0.94 (0.78 – 1.0)	0.91 (0.82 – 1.0)

*Indicate the best initial survival predictive ML models developed.

Table 6. Mean and range (in parathesis) of performance metrics for various refined ML survival models

Predictive Models	ML model	Accuracy (in %)	Precision	Recall	F1
Single model	**Random Forest***	**87.84 (80.95 – 95.45)**	**0.87 (0.81 – 0.95)**	**1.0**	**0.93 (0.89 – 0.97)**
	AdaBoost	86.72 (76.19 – 95.45)	0.87 (0.80 – 0.95)	0.98 (0.88 – 1.0)	0.92 (0.86 – 0.97)
	Decision Tree	82.43 (71.43 – 95.24)	0.86 (0.80 – 0.94)	0.93 (0.78 – 1.0)	0.89 (0.82 – 0.97)
	Logistic Regression	87.45 (77.27 – 95.45)	0.87 (0.81 – 0.95)	0.99 (0.88 – 1.0)	0.93 (0.86 – 0.97)
	Linear Support Vector Machine	84.86 (71.43 – 95.45)	0.88 (0.79 – 1.0)	0.94 (0.83 – 1.0)	0.91 (0.83 – 0.97)
Male model	Random Forest	89.93 (80.0 – 100.0)	0.89 (0.80 – 1.0)	1.0	0.94 (0.89 – 1.0)
	AdaBoost	88.47 (72.73 – 90.91)	0.88 (0.80 – 0.90)	0.99 (0.89 – 1.0)	0.93 (0.84 – 0.95)
	Decision Tree	83.34 (54.54 – 100.0)	0.90 (0.80 - 1.0)	0.89 (0.55 – 1.0)	0.89 (0.67 – 1.0)
	Logistic Regression	92.22 (80.00 – 100.0)	0.92 (0.80 – 1.0)	0.99 (0.89 – 1.0)	0.95 (0.89 – 1.0)
	Linear Support Vector Machine*†	**92.62 (81.82 – 100.0)**	**0.93 (0.82 – 1.0)**	**0.99 (0.89 – 1.0)**	**0.96 (0.89 – 1.0)**
Female model	Random Forest	90.47 (80.00 – 100.0)	0.90 (0.80 – 1.0)	1.0	0.95 (0.89 – 1.0)
	AdaBoost	90.80 (70.00 – 100.0)	0.92 (0.78 – 1.0)	0.99 (0.87 – 1.0)	0.95 (0.82 – 1.0)
	Decision Tree	81.38 (60.00 – 90.91)	0.88 (0.75 – 1.0)	0.90 (0.67 – 1.0)	0.88 (0.75 – 0.95)
	Logistic Regression*†	**91.96 (70.00 – 100.0)**	**0.93 (0.78 – 1.0)**	**0.99 (0.87 – 1.0)**	**0.95 (0.82 – 1.0)**
	Linear Support Vector Machine	91.20 (70.00 – 100.0)	0.93 (0.78 – 1.0)	0.98 (0.87 – 1.0)	0.95 (0.82 – 1.0)

*Indicate the best model of each group.

†Indicate the final models deployed to develop Gap-App service.

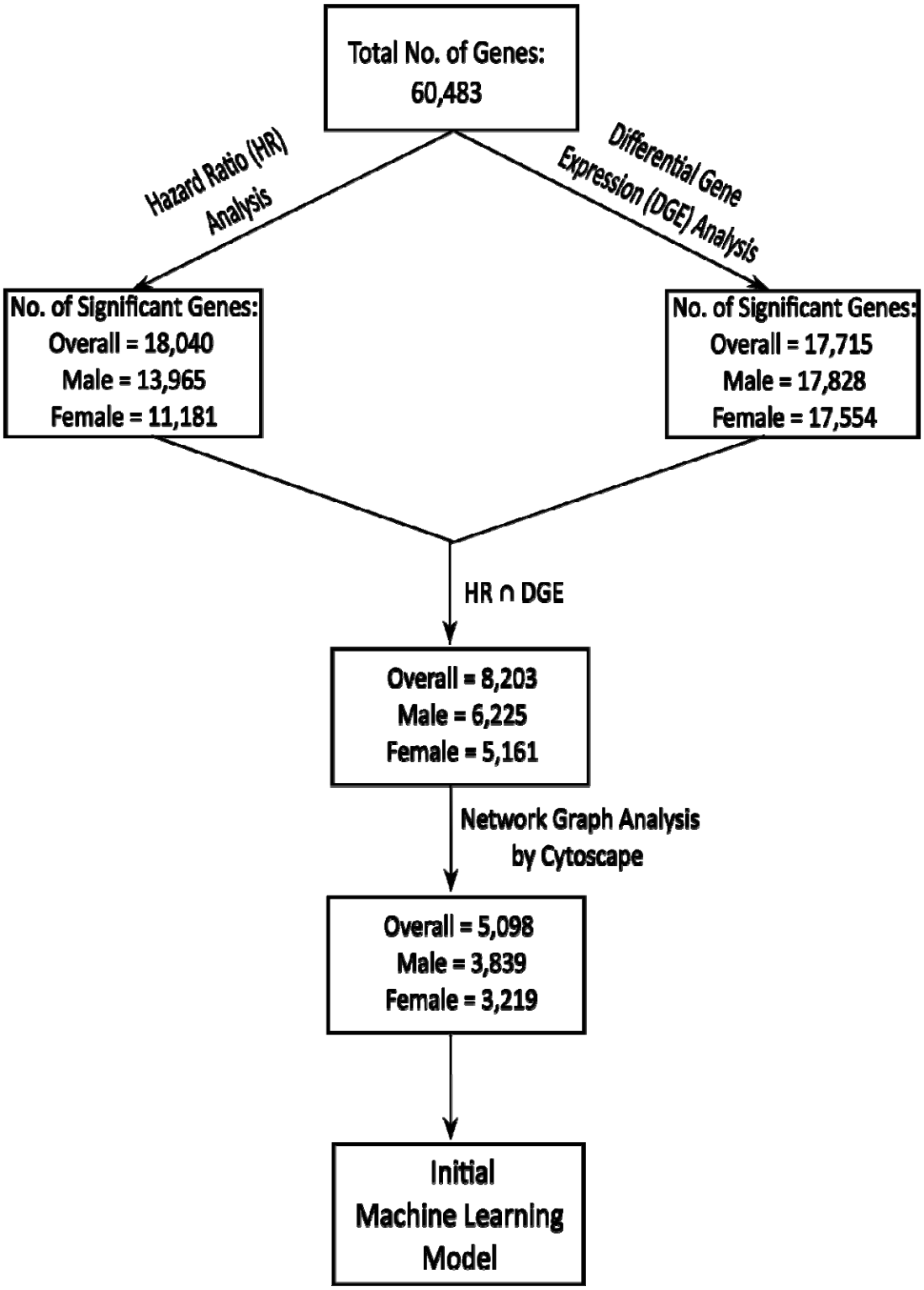

Figure 1. Flow chart of initial gene filtration.

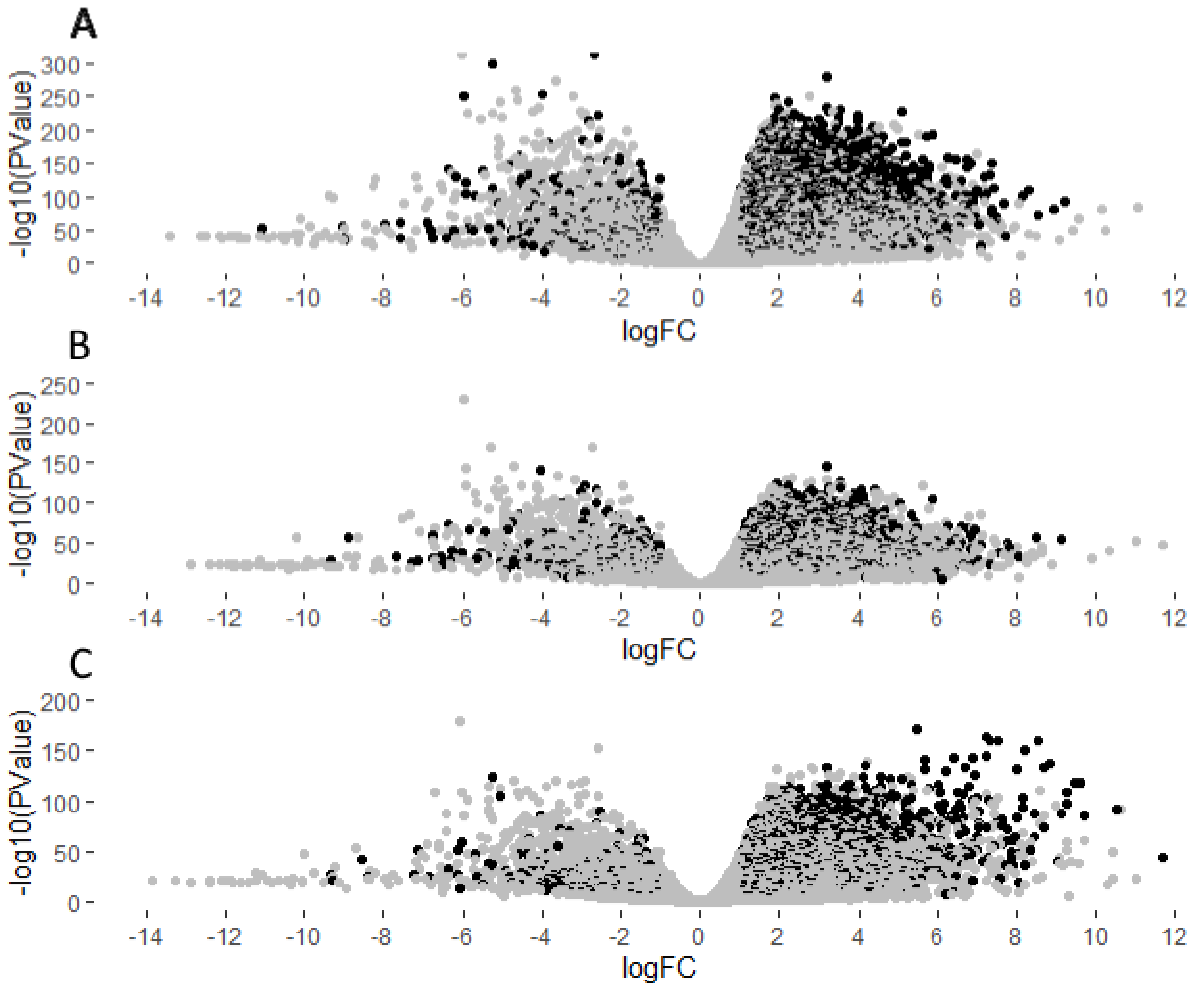

Figure 2: Volcano plot of DGE analysis for overall population (A), and male (B) and female (C) subgroups. Genes satisfying all three selection criteria for inclusion in machine learning models are highlighted in black, otherwise in grey.

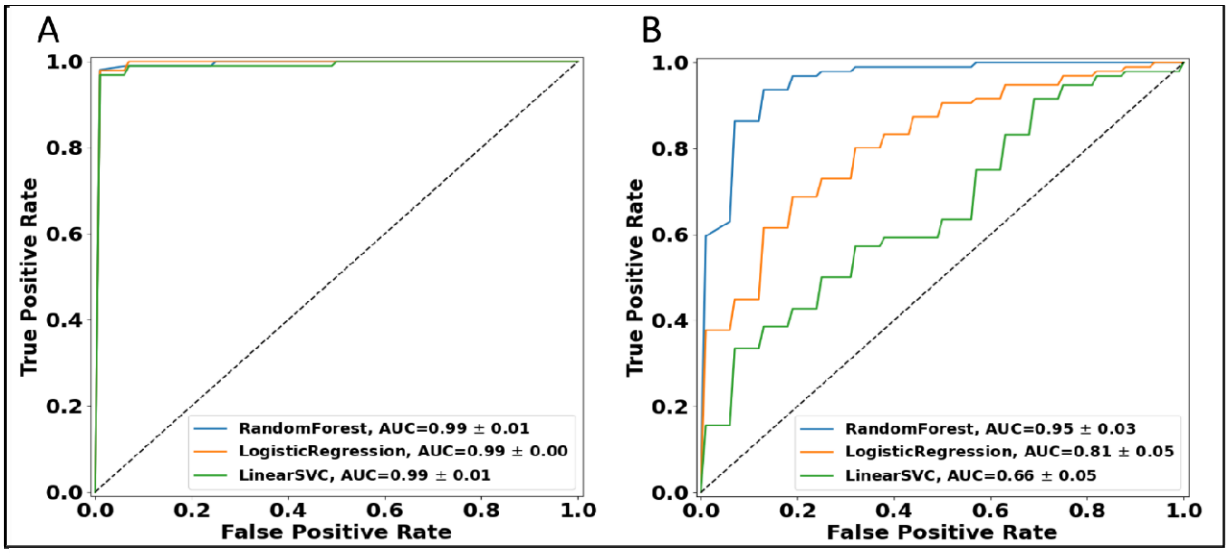

Figure 3: ROC curve for the sex predictive ML models with (A) or without (B) the Y-chromosome genes.

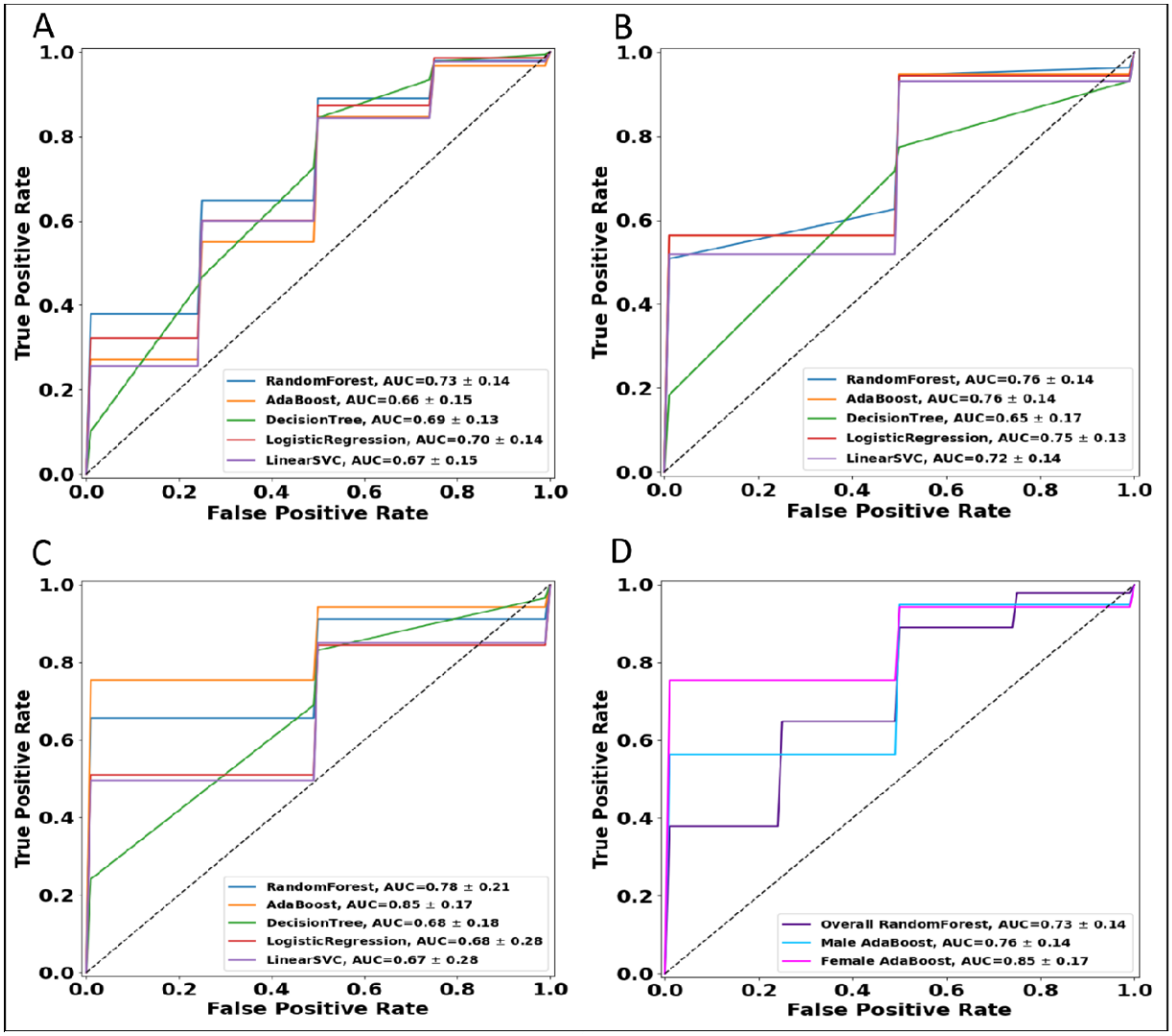

Figure 4. ROC curve of the initial 3-year survival predictive models using RNA-Seq reads as input features for overall (A), and male (B) and female (C) subgroups. ROC curve for the best models for overall (purple), male (blue) and female (pink) subgroups is shown in D.

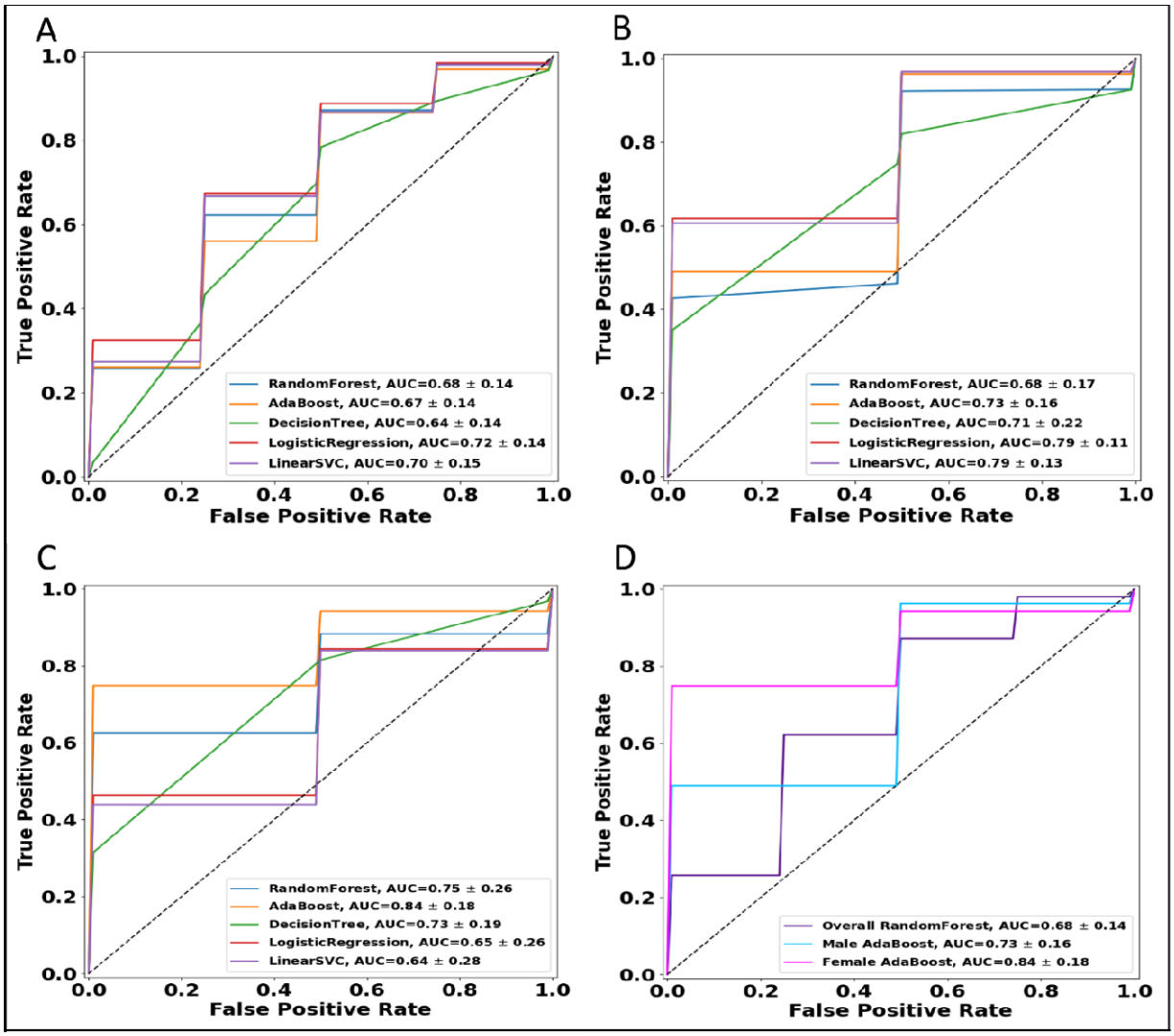

Figure 5. ROC curve of the initial 3-year survival predictive models using RNA-Seq reads and patients' characteristics as input features for overall (A), and male (B) and female (C) subgroups. ROC curve for the best models for overall (purple), male (blue) and female (pink) subgroups is shown in D.

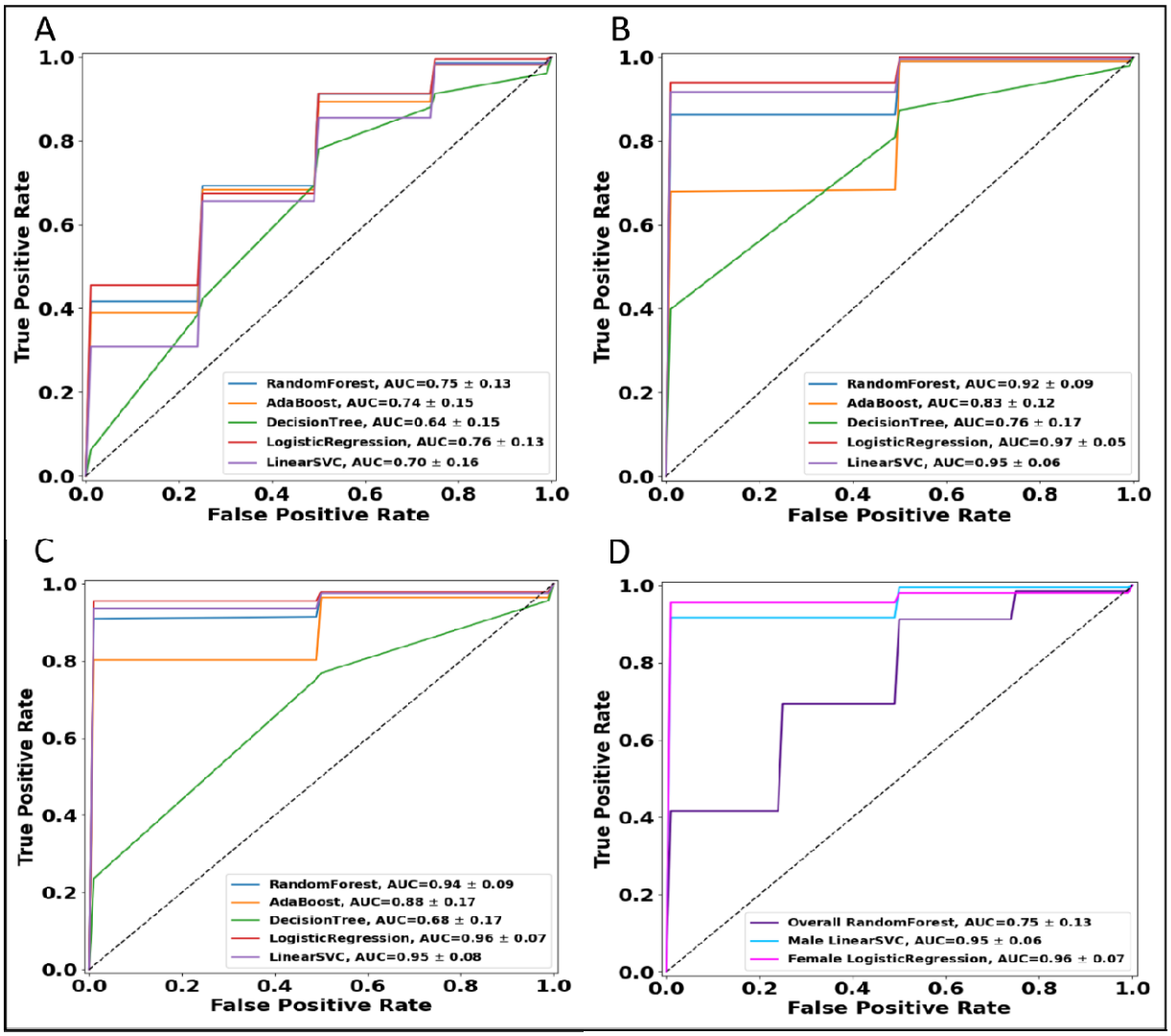

Figure 6. ROC curve of the refined 3-year survival predictive models using RNA-Seq reads as input features for overall (A), and male (B) and female (C) subgroups. ROC curve for the best models for overall (purple), male (blue) and female (pink) subgroups is shown in D.

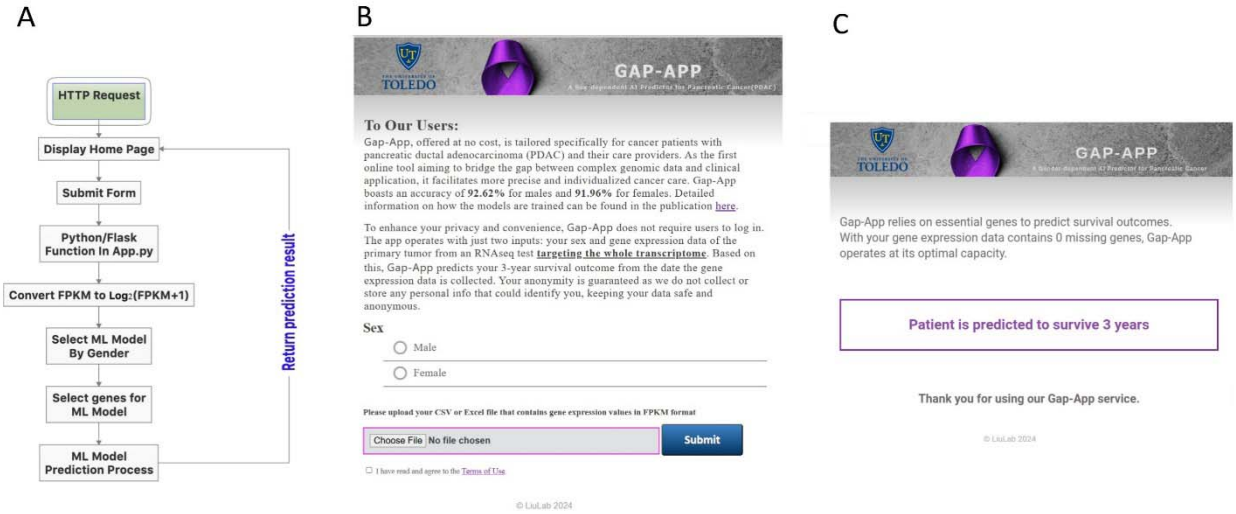

Figure 7: The developed Gap-App web application. (A) Workflow that outlines how Gap-App handles a prediction request. (B) Screenshot of Gap-App upon loading the webpage. (C) Screenshot of Gap-App after prediction is made.