IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 December 2024, accepted 17 December 2024, date of publication 19 December 2024,
date of current version 30 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3520550

RESEARCH ARTICLE

WeiFu: A Novel Pan-Cancer Driver Gene Identification Method Using Incidence-Weighted Mutation Scores

YANJIE REN[1,2], AZLAN MOHD ZAIN[1], (Member, IEEE), YAN ZHANG[2], ROZITA ABDUL JALIL[3], MAHADI BAHARI[4], NORFADZLAN BIN YUSUP[5], MAZLINA ABDUL MAJID[6], AZURAH A. SAMAH[1], DIDIK DWI PRASETYA[7], (Associate Member, IEEE), AND NURHAFIZAH MOZIYANA MOHD YUSOP[8]

[1]Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[2]Hebei Institute of Mechanical and Electrical Technology, Xingtai 054001, China
[3]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Malaysia
[4]Faculty of Management, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[5]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan 94300, Malaysia
[6]Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pahang 26600, Malaysia
[7]Faculty of Engineering, State University of Malang, Malang 65145, Indonesia
[8]Faculty of Defence Science and Technology, National Defence University of Malaysia, Kuala Lumpur 57000, Malaysia

Corresponding author: Yanjie Ren (renyanjie@graduate.utm.my)

This work was supported in part by the Ministry of Higher Education Malaysia (MOHE) under Fundamental Research Grant Scheme (FRGS) under Grant R.J130000.7809.5F524, and in part by the Key Project of Natural Science Foundation of Hebei Institute of Mechanical and Electrical Technology under Grant ZD24006.

ABSTRACT Genetic and genomic variations are primary drivers of tumor development. Identifying driver genes from numerous passenger genes across pan-cancer poses a significant challenge due to varying mutation loads. While independent studies have elucidated cancer-associated mutation patterns within specific cancer types, a systematic approach to integrating these mutation data for assessing the impact of gene mutations has been lacking. This study addresses this gap by integrating pan-cancer genomic somatic mutation data and introducing a novel mutation weight fusion (WeiFu) score calculation method. WeiFu computes frequency and weighted fusion scores by cancer type, facilitating the identification of potential driver genes. Evaluation results on an integrated pan-cancer dataset comprising 29 different cancer types demonstrate that WeiFu significantly outperforms current well-known approaches in prediction accuracy, sensitivity, and specificity. Notably, WeiFu recovers 277 known cancer genes among the top 500 ranked candidates and successfully identifies potential driver genes supported by strong evidence. Consequently, WeiFu shows considerable promise for identifying driver genes within the rapidly expanding corpus of cancer genomic data.

INDEX TERMS Driver gene, pan-cancer, somatic mutation, cancer incidence weighting.

I. INTRODUCTION

Cancer remains one of the major global health threats, characterized by its complexity, involving diverse genetic and epigenetic alterations. Over the past decade, advancements in high-throughput sequencing technologies have facilitated the detailed characterization of cancer genomes within individual cancer types. However, the heterogeneity and complexity of the disease necessitate a pan-cancer approach that transcends individual cancer types to elucidate broad

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

oncogenic mechanisms and identify potential universal driver genes. Identifying driver genes is crucial for diagnosis, prevention, and clinical treatment. The discovery of these pivotal driver genes enhances our understanding of tumorigenesis mechanisms and enables more effective identification of cancer pathways [1].

While traditional biological experiments have been instrumental in identifying of driver genes, they often require substantial time and financial investments. Consequently, the development of effective algorithms for driver gene identification through computational methods has become imperative. Somatic mutations have been the most extensively studied aspect of human carcinogenesis [2]. Somatic aberration data has been the predominant modality employed for driver gene identification, largely due to the accessibility of public databases such as The Cancer Genome Atlas (TCGA) [3], the Catalogue of Somatic Mutations in Cancer (COSMIC) [4], and the International Cancer Genome Consortium (ICGC) [5]. These public databases provide large-scale, high-quality somatic mutation data, which is essential for training and evaluating computational algorithms. In cancer genome projects, copy number variations (CNVs), single nucleotide polymorphisms (SNPs), and insertions/deletions (INDELs) are the three primary types of genetic variations that impact gene function and cancer progression.

A multitude of bioinformatics tools designed for driver gene identification using somatic mutation data have been developed. These tools can generally be categorized into three distinct groups based on their fundamental features. The first category focuses on frequency, where methods identify driver genes based on their higher mutation rates compared to the background mutation rate (BMR). Recurrence-based algorithms, such as MutSig2.0 [6], MutSiC [7], and MutSigCV [8], rank genes according to the likelihood of observing a given number of somatic single-nucleotide variants (SNVs) by chance. The second category centers on clustering, where mutations within particular regions often form clusters. Methods such as OncodriverCLUST [9] and HotMAPS [10] are designed to detect driver genes by identifying significant clusters of mutations in particular sequence regions. The third category is based on functional impact (FI). The primary challenge for these algorithms is assessing the effect of mutations. Typically, open-source tools are employed to address this challenge. For example, OMEN [11], MutationAssessor [12], and PolyPhen2 [13] assess mutation impact, whereas OncodriveFML [14] employs a scoring framework known as Combined Annotation Dependent Depletion (CADD) and RNAsnp to evaluate the impact of mutations across both coding and noncoding elements.

The significant variability in mutation burden across cancer types poses a challenge for driver gene identification methods that rely on mutation frequency calculated from entire cancer cohorts, potentially introducing biases [15]. To address this issue, we propose integrating CNV, SNP and INDEL data from 29 different cancers[1] to construct a comprehensive pan-cancer gene mutation matrix. By combining these diverse data types, we aim to capture a more complete picture of the genetic alterations across cancers. However, to further refine our approach and address the disparity in cancer prevalence, we also incorporate cancer incidence data. While previous studies have often overlooked the significance of cancer incidence data in driver gene identification, our approach explicitly incorporates this factor to address the disparity in cancer prevalence. By weighting driver genes based on cancer incidence, we can account for the fact that some cancers are more common than others, and therefore, driver genes associated with these cancers may be more important to identify. By incorporating these weighted cancer incidence data into our analysis, this study aims to identify driver genes through the calculation of a novel, composite mutation score for each gene, which reflects both mutation patterns and cancer prevalence. Our strategy comprises three main steps:

1) Pan-cancer somatic mutation matrix construction: We acquire CNV, SNP and INDEL data from curated public databases. After rigorous quality control, annotation, and filtering processes, we integrate these datasets into a binary matrix, excluding non-gene elements to focus on relevant genetic features.
2) Cancer type-specific frequency calculation: We stratify samples by cancer type and compute gene mutation frequencies within each group. This approach mitigates potential biases arising from inter-cancer type differences in mutation frequency, which can distort when calculating frequencies across heterogeneous samples.
3) Weighted fusion application: We assign weights to driver genes based on cancer incidence data from GLOBOCAN 2022 [16]. Several weight fusion approaches have achieved significant success [17], [18]. This weighting scheme reflects the real-world distribution of cancer types, enhancing the generalizability of our findings and reducing bias by ensuring adequate representation of common cancers. Finally, we merge these weighted scores to generate a comprehensive mutation score for each gene.

[1]The 29 different cancers include Adrenocortical Carcinoma (ACC), Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Colon Adenocarcinoma (COAD), Esophageal carcinoma (ESCA), Glioblastoma Multiforme (GBM), Head and Neck Squamous Cell Carcinoma (HNSC), Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney Renal Papillary Cell Carcinoma (KIRP), Brain Lower Grade Glioma (LGG), Liver Hepatocellular Carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Mesothelioma (MESO), Ovarian Serous Cystadenocarcinoma (OV), Pancreatic Adenocarcinoma (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate Adenocarcinoma (PRAD), Rectum Adenocarcinoma (READ), Sarcoma (SARC), Stomach Adenocarcinoma (STAD), Testicular Germ Cell Tumors (TGCT), Thyroid Cancer (THCA), Uterine Corpus Endometrial Carcinoma (UCEC), Uterine Carcinosarcoma (UCS), Uveal Melanoma (UVM).

II. MATERIALS AND METHODS

Our algorithm pipeline consists of four sequential steps (Figure 1):

1) Data acquisition and preprocessing: Somatic mutation data is collected from the TCGA Pan-Cancer database. This data undergoes preprocessing and is formatted into matrices where rows represent genes and columns represent samples.
2) Matrix integration: The diverse somatic mutation matrices are consolidated into a unified, comprehensive matrix.
3) Cancer-specific analysis: The integrated matrix is partitioned into separate cancer-specific somatic mutation matrices based on sample information. Gene mutation scores are then calculated for each cancer-specific matrix.
4) Weighted aggregation: Gene mutation scores are aggregated across cancers, applying weights derived from cancer incidence data.

A. DATA ACQUISITION AND PREPROCESSING

This study utilized publicly available datasets from The Cancer Genome Atlas (TCGA) database, accessible via the UCSC Xena Browser [19](https://xenabrowser.net/datapages/), covering 29 different cancer types. These datasets were selected for their comprehensive coverage of genomic alterations, including copy number variations (CNVs) in 24,777 genes across 10,845 samples, as well as single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) in 40,544 genes from 9,104 samples.[2] Detailed information is provided in Table 1.

A gene mutation matrix was constructed by integrating CNV, SNP and INDEL data across the 29 cancer types. In this matrix, rows represent genes, and columns correspond to samples, with binary entries indicating the presence (1) or absence (0) of a mutation in each gene-sample pair.

TCGA pan-cancer gene-level CNV estimates were calculated using the GISTIC2 threshold method, which compiled data from all TCGA cohorts. GISTIC2 categorized these estimates into five thresholds: $-2, -1, 0, 1$, and 2, as detailed in Table 2. Based on the interpretation of the estimated values, the matrix entries for each gene and sample are specifically defined as follows:

$$M_{CNV} = \begin{cases} 1, & M_{CNV} = 2 \quad or \quad M_{CNV} = -2 \\ 0, & otherwise \end{cases} \tag{1}$$

The TCGA pan-cancer dataset includes SNPs and INDELs from the Unified Ensemble ''MC3'' gene-level mutation calls. In the ''MC3'' dataset, each gene is classified based on mutation occurrence, where ''1'' indicating a non-silent mutation and ''0'' indicating a wild-type (wt) gene.

[2]The CNVs, SNPs, and INDELs data used in this study were pre-processed by the TCGA database prior to download, which involved quality control, alignment, and variant calling.

B. MATRIX INTEGRATION

In this study, we established a systematic approach for integrating multiple types of somatic mutation data. The fundamental element of our integration strategy is the standardization of mutation representation through a binary classification system, defined by Equation2:

$$M_{gs} = \begin{cases} 1, & M_{gs} = 2 \quad or \quad M_{gs} = 1 \\ 0, & otherwise \end{cases} \tag{2}$$

where M_{gs} represents the mutation status of gene g in sample s. This binary transformation enables the unified representation of mutations across different data types, with 1 indicating the presence of a mutation and 0 denoting its absence.

The integration procedure comprised three sequential phases. Initially, we performed a comprehensive alignment of samples and genes across three distinct mutation categories: CNVs, SNPs and INDELs. Following the alignment, we applied Equation 2 to standardize the mutation representations across all data types. The integration yielded an initial somatic mutation matrix encompassing 23,384 genes and 8,893 samples.

To enhance the biological relevance of our dataset, we implemented a systematic refinement protocol. This process involved the elimination of uninformative rows and columns (those containing exclusively zero values) and the careful exclusion of specific RNA entities: 1,254 miRNAs, 391 lncRNAs, and 1,662 ncRNAs. The refined matrix ultimately comprised 20,077 mutated genes across 8,891 samples, representing a comprehensive landscape of somatic mutations suitable for subsequent analyses.

C. CANCER-SPECIFIC ANALYSIS

To comprehensively account for multiple critical factors, we designed a novel scoring algorithm for identifying driver genes. In this algorithm, each cancer sample with gene mutations contributes equally to the overall score, reflecting research showing that only a limited number of driver mutation genes exist, irrespective of the overall count of mutated genes within a sample.

The algorithm begins by grouping samples according to cancer type and calculating the mutation frequency of each gene within each cancer group separately. By analyzing gene mutation frequencies within specific cancer groups, rather than across the entire mixed sample, this approach helps prevent the influence of mutation frequency differences among different cancer types. Figure 2 illustrates the mapping of 7,903 samples to 29 cancer types. We divided the mutation matrix into 29 sub-matrices based on samples corresponding to each cancer type and calculated the mutation score for each gene. Previous studies have shown that driver gene sets exhibit high coverage, meaning they are present in a significant number of samples. Therefore, we calculate a mutation score for each gene to quantify its contribution to cancer.

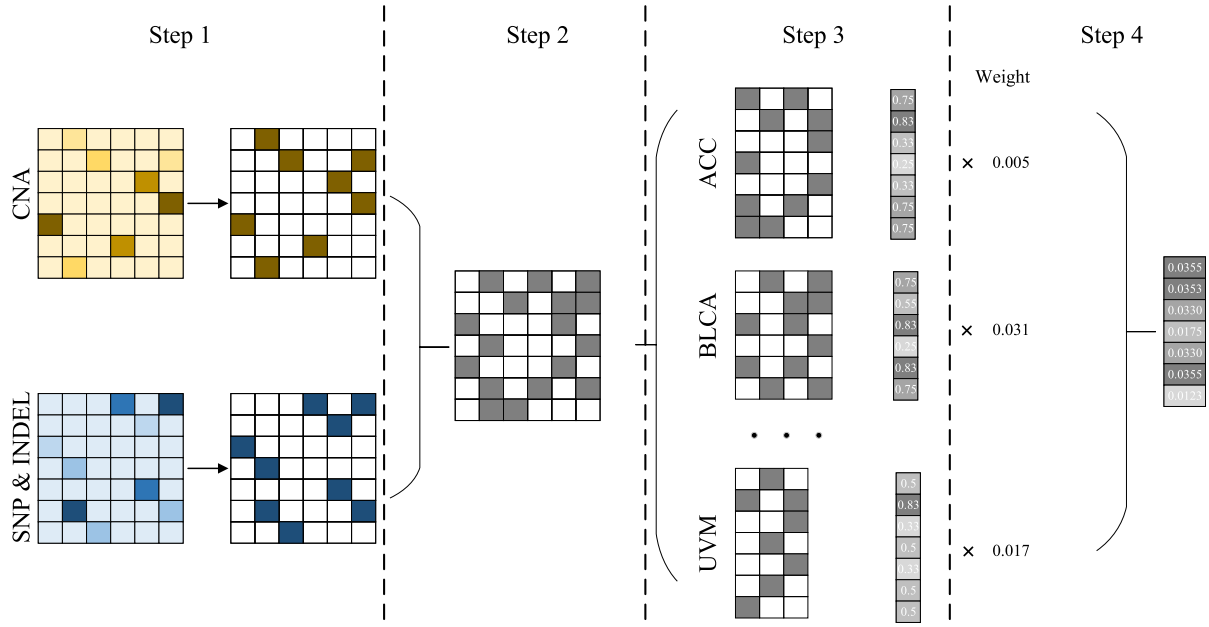

FIGURE 1. The WeiFu method for calculating somatic mutation scores employs a sophisticated four-step pipeline. Initially, somatic mutation data, including CNA, SNP and INDEL, are collected from public databases and preprocessed into binary matrices. These matrices are then integrated into a single unified matrix, providing a comprehensive view of genomic alterations. Subsequently, this unified matrix is partitioned into 29 submatrices, each representing a specific cancer type (e.g. ACC, BLCA, and UVM). For each cancer type, a unique weight is assigned (e.g., 0.005 for ACC, 0.031 for BLCA, 0.017 for UVM), and mutation scores are computed using a specific algorithm. Finally, these cancer-specific scores are merged, accounting for their respective weights, to calculate a comprehensive mutation score for each potential driver gene. This approach enables WeiFu to handle multiple types of mutations, account for cancer-type specificity, and provide a pan-cancer perspective on the likelihood of a gene being a driver, offering a nuanced and holistic understanding of somatic mutations in cancer genomics.

TABLE 1. Somatic mutation data sources.

Data Type	Samples	Identifiers	Size	Dataset
CNV	10845	24777	1.04 GB	https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.PANCAN.sampleMap/FGistic2_CopyNumber_Gistic2_ all_thresholded.by_genes.gz
SNP & INDEL	9104	40544	704 MB	https://gdc.cancer.gov/about-data/publications/mc3-2017

For a gene i, if there is a somatic mutation in a sample, then $M_{gs}(i,j) = 1$. Otherwise, $M_{gs}(i,j) = 0$. The mutation score $mu\_score(\cdot)$(Figure 3) for each gene i is calculated as follows:

$$mu\_score(i) = \begin{cases} \sum_{j \in \Gamma(i)} \frac{1}{|\Gamma(i)|}, & \Gamma(i) \neq \emptyset \\ \frac{1}{N_{max}}, & \Gamma(i) = \emptyset \end{cases} \quad (3)$$

where $\Gamma(i) = i : M_{ij} = 1$ represents the set of samples where gene i exhibits a mutation. $|\Gamma(i)|$ denotes the overall count of mutated genes in sample j, while the value of N_{max} is the maximum count of mutated genes across entire patients. When gene i does not exhibit a mutation in any patient, i.e. $\Gamma(i) = \emptyset$, setting the mutation score to $1/N_{max}$ helps identify driver genes with relatively low mutation rates. When $\Gamma(i) \neq \emptyset$, $1/|\Gamma(i)|$ quantifies the driving effect of the mutation in gene i for the disease state of the patient. The experiments were run on a system with an Intel 2.80-GHz i7-1165G7 CPU, 32GB of DDR4 RAM, and NVMe SSDs with a minimum of 1 TB storage.

If gene i is mutated in a sample, the mutation score of gene i is evenly distributed across all samples where it is mutated, irrespective of the overall mutated genes within that sample. Thus, the equation 3 balances contributions from all samples with varying numbers of mutated genes. In instances where gene i is absent from all mutated samples, its mutation score is assigned the background mutation score (BMS). Crucially, the BMS is set below the mutation score of any gene exhibiting at least one mutation. This ensures that the BMS helps identify potential driver genes characterized by exceptionally low mutation frequencies. This approach ensures that even genes with minimal mutation occurrences can be considered as potential drivers within the functional network.

D. WEIGHTED AGGREGATION

Given the significant variability in mutation burden across different cancer types, directly using gene mutation frequency calculated across the entire cancer cohort may introduce biases, potentially affecting the identification of driver

TABLE 2. CNVs estimated values.

Number	Represents	Meaning
-2	Homozygous deletion	Both copies of a particular genomic segment have been deleted
-1	A single copy deletion	One copy of the genomic segment has been deleted
0	A diploid normal copy	Both copies of the genomic segment are at the normal, diploid level
1	Low-level copy number amplification	A low increase in the number of copies of the genomic segment
2	High-level copy number amplification	A significant increase in the number of copies of the genomic segment

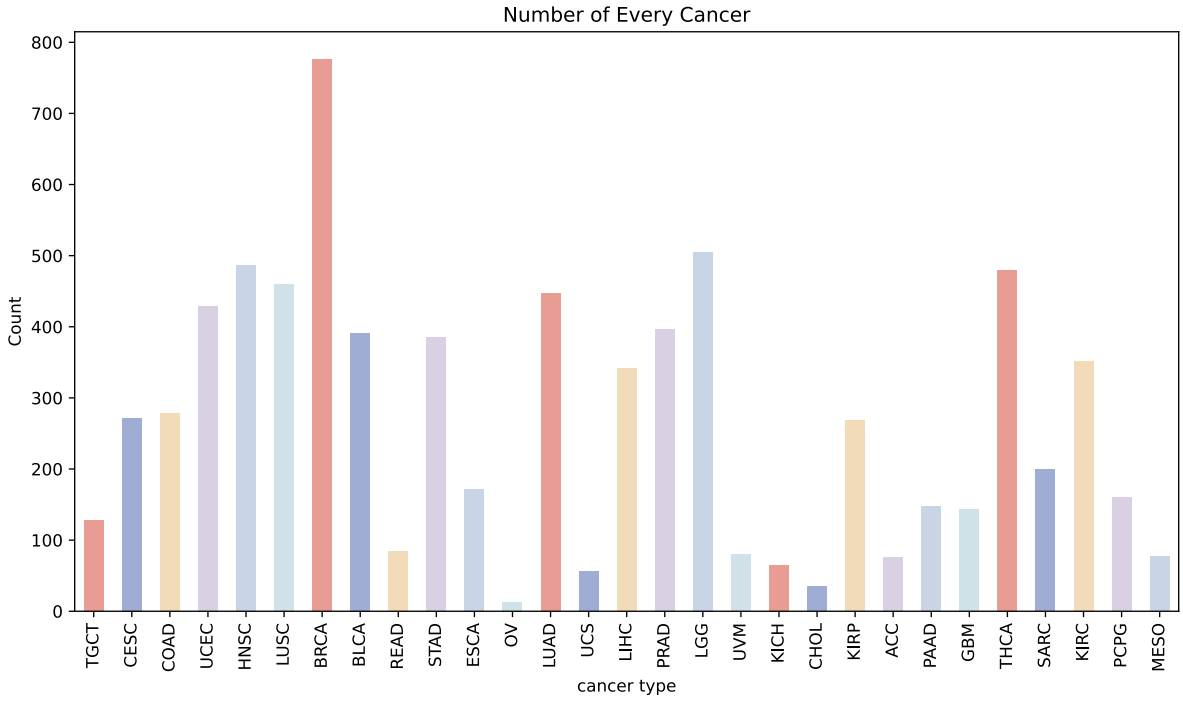

FIGURE 2. The dataset, containing the relationships between 7,903 samples and 29 cancer types, was downloaded from https://xenabrowser.net/.

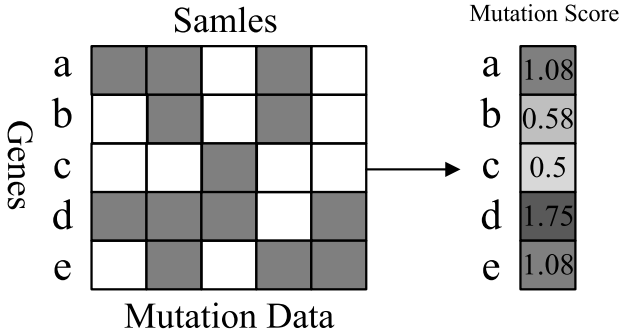

FIGURE 3. Illustration of mutation score. For example: $mu\_score(a) = 1/2 + 1/4 + 1/3 = 1.08$, $mu\_score(a) = 1/4 + 1/3 = 0.58$. The higher the score, the darker the colour of the gene.

TABLE 3. Cancer Weights.

Cancer	ACC	BLCA	BRCA	CESC	CHOL	COAD
weight	0.004	0.031	0.116	0.033	0.022	0.048
Cancer	ESCA	GBM	HNSC	KICH	KIRC	KIRP
weight	0.026	0.008	0.009	0.004	0.006	0.005
Cancer	LGG	LIHC	LUAD	LUSC	MESO	OV
weight	0.008	0.021	0.062	0.062	0.002	0.016
Cancer	PAAD	PCPG	PRAD	READ	SARC	STAD
weight	0.026	0.004	0.073	0.048	0.002	0.049
Cancer	TGCT	THCA	UCEC	UCS	UVM	
weight	0.004	0.041	0.011	0.010	0.017	

genes. We employed cancer incidence as a weighting factor(Table 3). This approach prioritizes interventions targeting cancers with higher incidence, reflecting the real-world distribution of cancer types and enhancing generalizability. Weighting by incidence reduces bias by ensuring that more common cancers are not underrepresented, leading to more balanced predictions. The latest incidence statistics for different cancer types were obtained from GLOBOCAN 2022 [16] and converted into incidence weights, as shown in Table 3. For example, if we consider a gene with a mutation score of 0.5 and an incidence weight of 0.116 (for BRCA), the weighted score would be calculated as 0.5×0.116. Ultimately, we developed a weighted integration strategy to combine driver gene scores across all cancer types, resulting in a composite scoring formula for identifying pan-cancer

driver genes. This ensures that more common cancers are adequately represented, reducing bias and leading to balanced predictions. We will provide visual aids, such as flowcharts, to illustrate these steps clearly. Ultimately, our composite scoring formula is defined as follows:

$$ScoreFinal(g) = \sum_{k=1}^{K} w(k) \times mu\_score(g) \tag{4}$$

Consider genes A and B. Gene A has a high mutation frequency in lung cancer (weight: 0.062) but a low mutation frequency in other cancers. Gene B has a moderate mutation frequency across multiple cancer types. After applying the weighting process, Gene B may have a higher overall score due to its consistent presence in various cancers, even though its mutation frequency in any individual cancer type might be lower than Gene A. We rank all genes based on the value of $ScoreFinal(g)$ and select the top-ranked genes as the predicted candidate set of pan-cancer driver genes.

III. RESULTS AND DISCUSSION

A. EVALUATION CRITERIA AND REFERENCE CANCER GENE SETS

To evaluate the performance of candidate driver gene ranking methods, we employed several criteria:

1) ROC curve analysis and AUC values to measure coverage of known driver genes
2) F1 score and cumulative number of known driver genes within the top 500 ranked candidates to assess the method's ability to concentrate true driver genes at the top of the ranking

Given the limited resources for experimental validation, we primarily focused on the top 500 ranked candidates. The F1 score, as the harmonic mean of precision and recall, offers a comprehensive reflection of classification performance. To eliminate potential bias from varying numbers of known driver genes recovered within the top 500 by different methods, recall was calculated based on the overall count of known driver genes within each reference cancer gene set. Both ROC curve analysis and statistical tests were conducted using the ''pROC'' R package [20], with the null hypothesis being no significant difference in AUC values between different methods.

To objectively evaluate methods designed to identify candidate driver genes, an ideal assessment would rely on an unbiased and comprehensive set of known cancer genes. However, such a definitive ''gold standard'' gene set does not currently exist. Therefore, this study collected three different cancer gene sets to minimize the bias introduced by using a single reference gene set [21]: (i) 748 genes from the Cancer Gene Census (CGC) (v100) - TIER1/2 data from COSMIC [4], the most widely used cancer gene set; (ii) 5025 genes from CancerMine [22], a literature-mined database of drivers, oncogenes and tumor suppressors in cancer; (iii) 1151 genes from OncoKB [23]. Despite the inherent trade-offs in terms of accuracy, reliability, comprehensiveness, and unbiasedness among the reference gene sets, an effective method should consistently demonstrate superior performance compared to other methods. WeiFu recovered 277 known cancer genes within the top 500 ranked candidates, demonstrating its efficacy in discerning potential cancer driver genes supported by substantial experimental validation.

B. CANCER ASSOCIATIONS

Each target in the query set was annotated with known associations to cancer phenotypes (ontology terms) as established from various data types through the Open Targets Platform (OTP v2024.06) [24]. Annotations are provided using both descriptive associations and rank scores. The identification of putative cancer driver genes required support from at least two of the following sources:Network of Cancer Genes (NCG) [25], IntOGen [26], Cancer Gene Census (CGC) (v100) - TIER1/2 data from COSMIC [4], and predicted cancer driver genes from TCGA [27] were utilized in our analysis.

We employed Galaxy Oncotools [28] for exploratory analysis and prioritization of the top 20 gene lists from high-throughput cancer biology experiments. As illustrated in Figure 4, certain genes, such as TP53, APC, and KRAS, exhibit strong and consistent associations across multiple cancer types, underscoring their fundamental roles in cancer biology. Conversely, other genes display cancer-specific associations, highlighting the significance of context in cancer genetics. Additionally, we provide a more detailed explanation of the molecular biology and functional implications of the top driver genes and their associations with various cancers (Table 4).

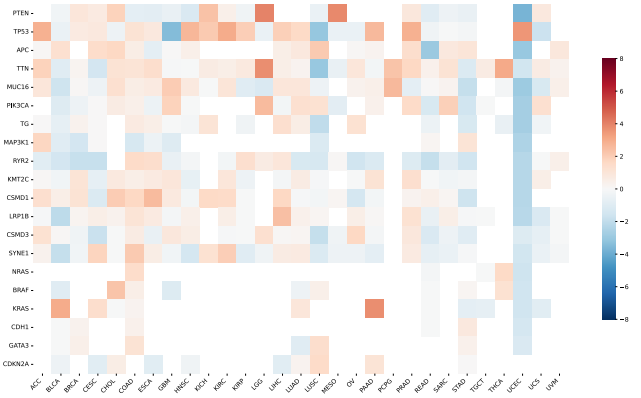

FIGURE 4. The results of gene-cancer associations. The presence of both positive (red) and negative (blue) associations indicates that mutations can have varied effects on survival depending on the gene and cancer type.

C. ENRICHMENT ANALYSIS OF CANCER-RELATED GENES

Building upon these cancer-specific gene associations, we further investigated the functional implications of the top-score-ranked genes through comprehensive enrichment

TABLE 4. The Explanation of The Molecular Biology of The Top Driver Genes and Their Associations with Various Cancers.

Gene	Associations	Cell Biology
TP53	Exhibits strong positive associations (blue) across many cancer types, indicating that TP53 mutations are generally linked to poor survival outcomes, consistent with its critical role as a tumor suppressor gene.	A tumor suppressor gene that regulates cell cycle, apoptosis, and DNA repair. Mutations can lead to uncontrolled cell proliferation and are prevalent in lung, breast, and colorectal cancers.
TTN	Demonstrates a mix of positive and negative associations, with a particularly notable strong positive association in LUAD (Lung Adenocarcinoma).	Encodes titin, important for muscle contraction and cellular mechanosensation. Mutations in LUAD may impact cellular architecture and tissue homeostasis.
CSMD1	Primarily shows weak associations, with a strong negative association (red) in UVM, suggesting a potential protective effect of mutations in this cancer type.	Negative regulator of cell adhesion and migration. Loss or mutation may enhance cellular adhesion or decrease migration, potentially slowing metastasis in UVM.
APC	Displays strong positive associations in COAD and READ (Colorectal cancers), aligning with its established role in colorectal cancer development.	Regulates the Wnt/β-catenin pathway. Mutations lead to β-catenin stabilization, promoting cell proliferation and survival in colorectal cancer.
KRAS	Shows strong positive associations in PAAD (Pancreatic Adenocarcinoma) and LUAD, consistent with its known oncogenic role in these cancers.	Encodes a GTPase involved in cell growth regulation. Mutations lead to constitutive activation, promoting oncogenesis in PAAD and LUAD.
PIK3CA	Displays mostly weak associations, with some positive correlations in breast cancers (BRCA).	Encodes a PI3-kinase subunit. Mutations activate the PI3K/AKT/mTOR pathway, promoting tumor growth and survival, particularly in breast cancer.
KMT2C	Shows mixed associations, with a notable positive correlations in KIRC (Kidney Renal Clear Cell Carcinoma).	Involved in histone modification and chromatin remodeling. Mutations can alter gene expression profiles, contributing to tumorigenesis in KIRC.
PTEN	Demonstrates strong positive associations in several cancer types, including UCEC (Uterine Corpus Endometrial Carcinoma), highlighting its critical role as a tumor suppressor.	A tumor suppressor that inhibits the PI3K/AKT pathway. Loss of function leads to pathway hyperactivation, promoting oncogenesis.
GATA3	Displays a strong negative association in BRCA, suggesting that mutations might correlate with better survival in breast cancer.	A transcription factor that regulates genes involved in cell differentiation. Negative association in BRCA suggests a protective role in breast cancer.
NRAS	Displays generally weak associations across various cancer types.	A GTPase involved in cell signaling. Its role in tumorigenesis is less clear compared to other RAS family members.
BRAF	Has a strong positive association in SKCM (Skin Cutaneous Melanoma)	A kinase in the MAPK pathway. Mutations lead to pathway activation, promoting tumor cell proliferation in melanoma.
CDH1	Shows a strong negative association in STAD (Stomach Adenocarcinoma), implying a potential protective effect.	Encodes E-cadherin, important for cell adhesion and preventing EMT. Loss of function promotes metastasis in STAD.
CDKN2A	Has notable positive associations in several cancer types, including HNSC (Head and Neck Squamous Cell Carcinoma).	Encodes a tumor suppressor that inhibits cyclin-dependent kinases. Mutations lead to uncontrolled cell division and are associated with poor prognosis in HNSC.

analysis. By employing multiple analytical frameworks—Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO) and Hallmark pathway collections—the research illuminates the complex genetic landscape underlying these malignancies. We performed enrichment analysis on the top 20 genes for each cancer type. Results for lung cancer (LUAD and LUSC) and breast cancer (BRCA), which exhibited the most significant impact, are displayed in Figure 5. Notably, neurological and cellular signaling pathways appear as recurring themes across different cancer types, suggesting potential pan-cancer mechanisms of genetic dysregulation. Each analysis also highlights pathways specific to the cancer type, such as breast cancer-specific pathways for BRCA and lung cancer-specific pathways for LUAD and LUSC.

The functional annotation of driver genes reveals distinct molecular signatures that underscore their potential oncogenic mechanisms. In BRCA, the significant enrichment of terms related to neurological development indicates that driver genes might engage with essential cellular pathways beyond the scope of conventional oncogenic processes. This novel molecular context suggests that the identification of driver genes should not be limited to mutation or expression analysis alone but should also encompass a broader range of functional insights to uncover their transformative potential.

Lung cancer subtypes display significant molecular distinctions. Driver genes in LUAD are notably enriched in contractile and calcium-mediated processes, suggesting potential roles in cellular plasticity and metastatic potential. In contrast, LUSC exhibits enrichment in neurotransmitter and ion channel activities, underscoring the importance of subtype-specific molecular characterization in the identification of driver genes. These findings emphasize that driver genes are not merely passive genetic elements but are active molecular regulators with context-specific functional implications.

D. PROTEIN PROTEIN INTERACTION NETWORK ANALYSIS

The PPI network was generated using STRING 12.0 [29], a database of known and predicted protein-protein interactions, including both direct (physical) and indirect (functional) associations. This analysis focused on the top 20 genes identified by the WeiFu method. As shown in Figure 6, remarkably, 17 out of the 20 identified genes contribute to the network, suggesting that most of the identified genes are functionally related at the protein level. The network contains 64 edges, indicating numerous interactions among these proteins, which points to their involvement in related biological processes. The extremely

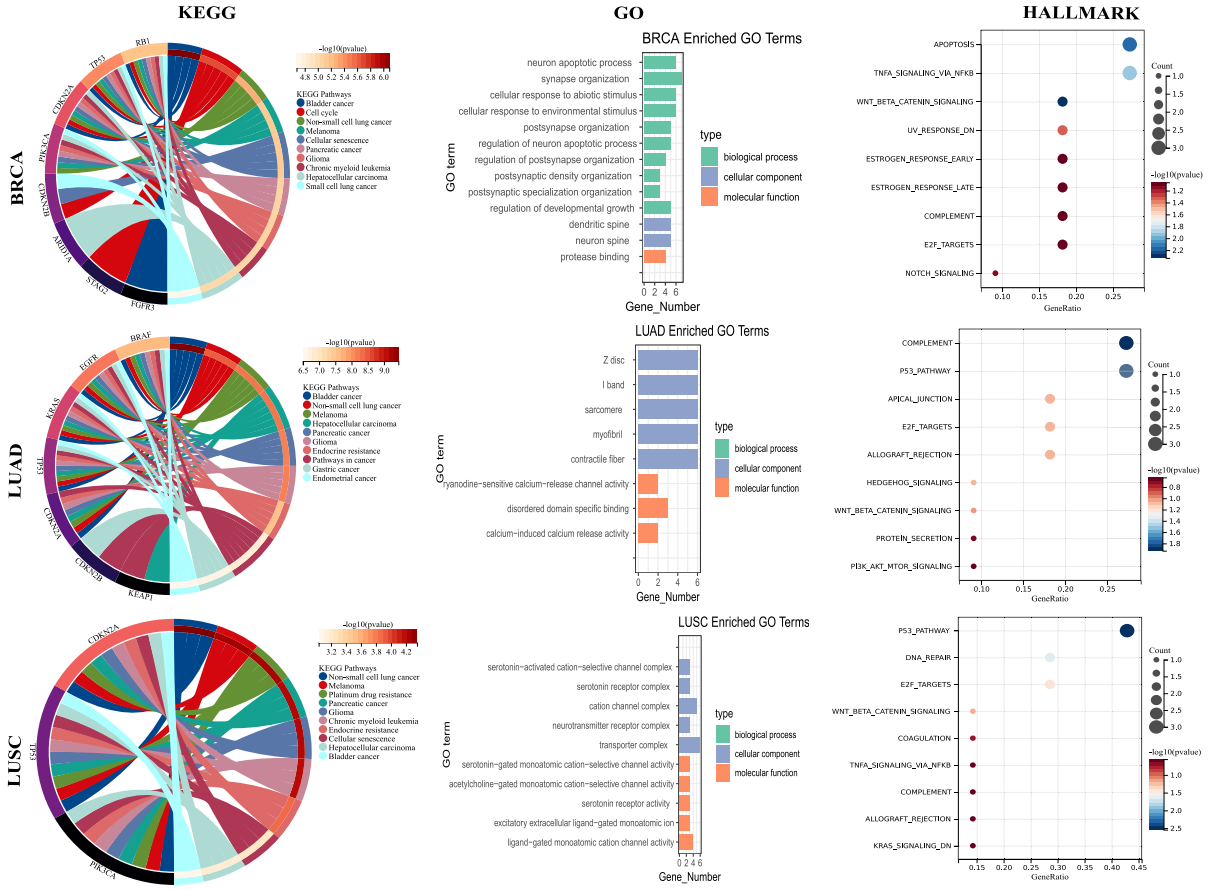

FIGURE 5. The results of driver gene enrichment. In the BRCA, LUAD and LUSC datasets, several significant overlapping biological processes were identified, including transcriptional activation activity, central carbon metabolism of cancer cells, apoptosis, and other pathways enriched in target diseases and cancer-related pathways.

low p-value (1.0E-16) indicates that there are significantly more interactions between these proteins than would be expected by chance for a random set of proteins of the same size. This strongly suggests that these proteins are biologically related and likely function in related pathways or processes. The presence of well-known cancer genes within this network further validates the effectiveness of the WeiFu method in identifying relevant cancer-associated genes.

E. METHODS COMPARISON

We evaluated the ability of WeiFu to differentiate between driver genes and passenger genes using our integrated pan-cancer dataset. Comparative analysis with other algorithms, including OMEN, ROI-Driver [30], MaxCLK [31], MutSigCV, and MTGCN [32], revealed that WeiFu significantly outperformed these methods in terms of both sensitivity and specificity. This superiority is evidenced by the ROC curves and AUC scores, as demonstrated in Figure 7.

The ROC curves comparison (7a) highlights the relative performance of various methods in balancing true positive rate and false positive rate across different thresholds. WeiFu emerges as the top-performing method, with its curve consistently above the others, closely followed by OMEN. These two methods demonstrate superior ability to distinguish between true positives and false positives compared to the rest. The remaining methods show a more varied performance. ROI-Driver, MaxCLK, and MutSigCV form a cluster with similar performance, their curves lying below WeiFu and OMEN but still showing improvement over random chance (represented by the diagonal line). While all methods demonstrate predictive power to varying degrees, there's a clear hierarchy in their performance. WeiFu and OMEN stand out as the most effective, with WeiFu holding a slight edge. The middle group of ROI-Driver, MaxCLK, and MutSigCV offer comparable performance to each other, while MTGCN lags behind. This analysis suggests that WeiFu would be the preferred choice for tasks requiring high discrimination between true and false positives, with OMEN as a strong alternative. The choice among ROI-Driver, MaxCLK, and MutSigCV might depend on factors beyond their ROC performance, given their similar curves.

The F1 Score Comparison graph (7b) illustrates the performance of six different methods for identifying cancer driver genes across varying numbers of genes (100 to 1,000). WeiFu consistently outperforms all other methods,

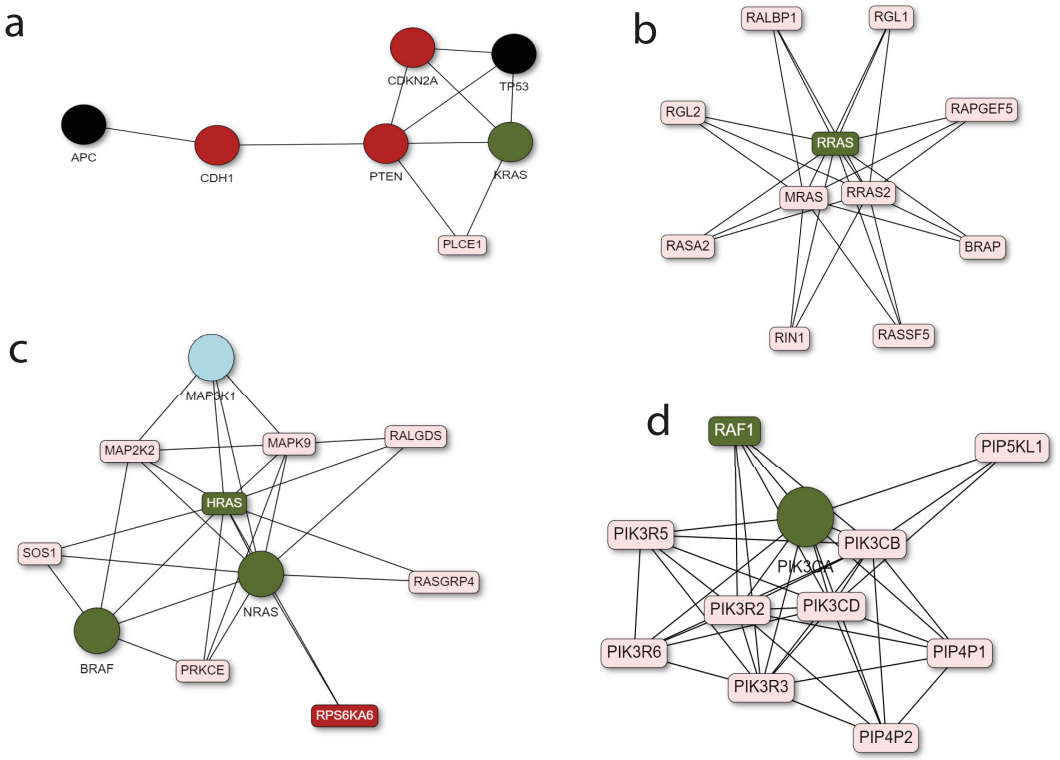

FIGURE 6. In the PPI network diagram of the top-20 ranked genes, each node represents a protein, and each edge represents an interaction. Of the 20 selected proteins, 17 contributed to the predicted PPI map with 64 edges and a PPI enrichment p-value of 1.0E-16. (a) Protein-containing complex binding (b) Cell differentiation (c) Central Role of RAS Family (d) PI3K Pathway.

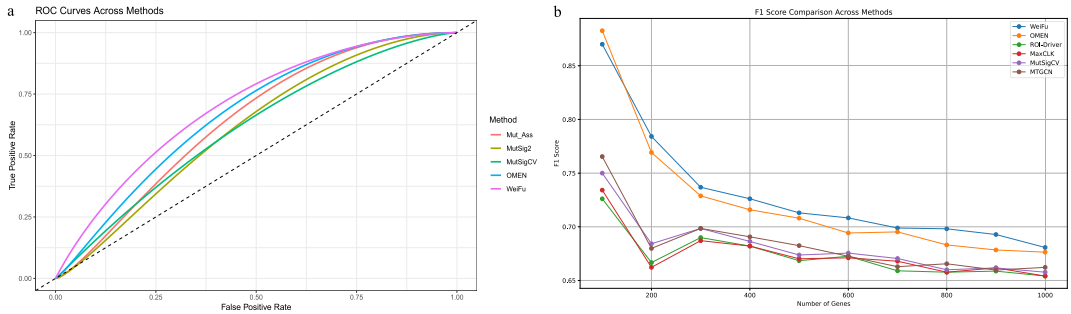

FIGURE 7. Thecomparative analysis of multiple methods for identifying driver genes. ROC curves and AUC scores, F1 Score Comparison Across Methods. The ROC curves show WeiFu's ability to achieve high sensitivity and specificity, while its F1 score underlines its superior balance in predicting true driver mutations.

maintaining the highest F1 score throughout the range, followed closely by OMEN. Both show a steady upward trend as the number of genes increases, with WeiFu reaching an F1 score of approximately 0.85 at 1,000 genes. ROI-Driver ranks third, while MaxCLK, MutSigCV, and MTGCN form a lower-performing group with more fluctuating scores. All methods demonstrate improved performance with larger gene sets, as evidenced by the general upward trend of their respective curves. The graph clearly demonstrates WeiFu's superiority in balancing precision and recall across different gene set sizes, with OMEN emerging as a strong second choice. This comparison provides valuable insights for researchers selecting methods for cancer driver gene identification, highlighting the effectiveness of WeiFu and OMEN, especially when analyzing larger gene sets.

IV. LIMITATION AND FUTURE WORK

While WeiFu demonstrates significant improvements in driver gene identification, several limitations require further consideration. Primarily, the study primarily relies on somatic mutation data, potentially overlooking other valuable multi-omic data, such as transcriptomics, proteomics, and

epigenomics. The integration of multi-omic data would provide a more comprehensive view of cellular processes and their dysregulation in cancer. Additionally, the current version of WeiFu may not fully capture the complex interactions between genes or the effects of mutations on protein structure and function.

Future research should focus on enhancing WeiFu by integrating multi-omic data. Integration of transcriptomic, proteomic, and other omic data types would provide additional insights into the functional implications of these mutations and their impact on cellular processes. Experimental validation of WeiFu's predictions, particularly for novel candidate driver genes, is essential to fully substantiate the computational findings. Future efforts should focus on collaborating with experimental or clinical researchers to validate these predictions in biological systems, which will provide a stronger foundation for real-world applicability. To enhance WeiFu's interpretability, we should incorporate advanced machine learning techniques like explainable AI (XAI) frameworks, such as SHAP and LIME, to clarify how genetic mutations interact. Additionally, employing ensemble learning methods like Random Forests can capture complex, non-linear relationships while providing insights through variable importance measures. Moreover, developing a user-friendly interface would promote broader adoption by the wider research community. These improvements would lead to a more comprehensive understanding of cancer driver genes and potentially inform targeted therapeutic strategies.

V. CONCLUSION

Driver genes play a crucial role in enabling accurate cancer diagnosis, accelerate the development of targeted drugs, and enable personalized treatment for clinical cases. However, distinguishing driver mutations from passenger mutations is challenging due to the fact that a small subset of genes exhibit mutations across a broad spectrum of cancers, while a substantial number of other genes are associated with specific cancer types. Although many existing methods prioritize differential mutation frequencies, they often demonstrate suboptimal sensitivity and specificity, as mutation frequency does not necessarily correlate with driver status. To address this, the WeiFu method incorporates several strategies:

1) Grouping samples by cancer type and calculating gene mutation frequency within each group
2) Assigning equal total weight to all mutated genes within each sample
3) Prioritizing candidate genes with lower mutation frequencies
4) Assigning different weights to driver genes in different cancers based on cancer incidence data from GLOBOCAN 2022 [16]

In a comparison with four other methods using pan-cancer data, WeiFu proves to be the most effective method overall. However, the performance of all methods declines as the number of genes increases, indicating that these tools are most reliable for identifying smaller sets of high-confidence gene candidates. WeiFu stands out as a valuable tool for the identification and prioritization of cancer driver genes, particularly in the context of the rapidly expanding landscape of cancer genomics.

The accurate identification of driver genes, as facilitated by methods like WeiFu, can significantly advance cancer research. This includes supporting early and accurate diagnosis, accelerating the development of targeted drugs by identifying actionable mutations, and paving the way for personalized treatment strategies tailored to patients' unique genetic profiles. The codes and datasets are available at https://github.com/Batelu/WeiFu.git.

REFERENCES

[1] Y. Kan, L. Jiang, J. Tang, Y. Guo, and F. Guo, "A systematic view of computational methods for identifying driver genes based on somatic mutation data," *Briefings Funct. Genomics*, vol. 20, no. 5, pp. 333–343, Sep. 2021.

[2] C. Swanton et al., "Embracing cancer complexity: Hallmarks of systemic disease," *Cell*, vol. 187, no. 7, pp. 1589–1616, Mar. 2024.

[3] C. G. A. R. Network, "Integrated genomic and molecular characterization of cervical cancer," *Nature*, vol. 543, no. 7645, p. 378, Apr. 2017.

[4] Z. Sondka et al., "COSMIC: A curated database of somatic variants and clinical data for cancer," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1210–D1217, Jan. 2024.

[5] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L. D. Stein, and V. Ferretti, "The international cancer genome consortium data portal," *Nature Biotechnol.*, vol. 37, no. 4, pp. 367–369, Apr. 2019.

[6] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz, "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature*, vol. 505, no. 7484, pp. 495–501, Jan. 2014.

[7] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, "MuSiC: Identifying mutational significance in cancer genomes," *Genome Res.*, vol. 22, no. 8, pp. 1589–1598, Aug. 2012.

[8] M. S. Lawrence, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, Jun. 2013.

[9] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes," *Bioinformatics*, vol. 29, no. 18, pp. 2238–2244, Sep. 2013.

[10] C. Tokheim, R. Bhattacharya, N. Niknafs, D. M. Gygax, R. Kim, M. Ryan, D. L. Masica, and R. Karchin, "Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure," *Cancer Res.*, vol. 76, no. 13, pp. 3719–3731, Jul. 2016.

[11] D. Van Daele, B. Weytjens, L. De Raedt, and K. Marchal, "OMEN: Network-based driver gene identification using mutual exclusivity," *Bioinformatics*, vol. 38, no. 12, pp. 3245–3251, Jun. 2022.

[12] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010.

[13] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: Application to cancer genomics," *Nucleic Acids Res.*, vol. 39, no. 17, p. e118, Sep. 2011.

[14] L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas, "OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations," *Genome Biol.*, vol. 17, no. 1, pp. 1–13, Dec. 2016.

[15] A. Gonzalez-Perez and N. Lopez-Bigas, "Functional impact bias reveals cancer drivers," *Nucleic Acids Res.*, vol. 40, no. 21, p. e169, Nov. 2012.

[16] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024.

[17] S. Anwar and N. Barnes, "Densely residual Laplacian super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1192–1204, Mar. 2022.
[18] Y. Yu, K. She, K. Shi, X. Cai, O.-M. Kwon, and Y. Soh, "Analysis of medical images super-resolution via a wavelet pyramid recursive neural network constrained by wavelet energy entropy," *Neural Netw.*, vol. 178, Oct. 2024, Art. no. 106460.
[19] A. S. Hinrichs, "The UCSC genome browser database: Update 2006," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D590–D598, Jan. 2006.
[20] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Muller, "PROC: An open-source package for r and S+ to analyze and compare ROC curves," *BMC Bioinf.*, vol. 12, no. 1, pp. 1–8, Dec. 2011.
[21] Y. Hou, B. Gao, G. Li, and Z. Su, "MaxMIF: A new method for identifying cancer driver genes through effective data integration," *Adv. Sci.*, vol. 5, no. 9, Sep. 2018, Art. no. 1800640.
[22] J. Lever, E. Y. Zhao, J. Grewal, M. R. Jones, and S. J. M. Jones, "CancerMine: A literature-mined resource for drivers, oncogenes and tumor suppressors in cancer," *Nature Methods*, vol. 16, no. 6, pp. 505–507, Jun. 2019.
[23] S. P. Suehnholz, "Quantifying the expanding landscape of clinical actionability for patients with cancer," *Cancer Discovery*, vol. 14, no. 1, pp. 49–65, Oct. 2023.
[24] D. Carvalho-Silva, A. Pierleoni, M. Pignatelli, C. Ong, L. Fumis, N. Karamanis, M. Carmona, A. Faulconbridge, A. Hercules, E. McAuley, A. Miranda, G. Peat, M. Spitzer, J. Barrett, D. G. Hulcoop, E. Papa, G. Koscielny, and I. Dunham, "Open targets platform: New developments and updates two years on," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1056–D1065, Jan. 2019.
[25] L. Dressler, M. Bortolomeazzi, M. R. Keddar, H. Misetic, G. Sartini, A. Acha-Sagredo, L. Montorsi, N. Wijewardhane, D. Repana, J. Nulsen, J. Goldman, M. Pollitt, P. Davis, A. Strange, K. Ambrose, and F. D. Ciccarelli, "Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: An update of the network of cancer genes (NCG) resource," *Genome Biol.*, vol. 23, no. 1, p. 35, Dec. 2022.
[26] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, and N. Lopez-Bigas, "A compendium of mutational cancer driver genes," *Nature Rev. Cancer*, vol. 20, no. 10, pp. 555–572, Oct. 2020.
[27] M. H. Bailey, "Comprehensive characterization of cancer driver genes and mutations," *Cell*, vol. 173, no. 2, pp. 371–385, Apr. 2018.
[28] (2024). *The Galaxy Platform for Accessible, Reproducible, and Collaborative Data Analyses: 2024 Update*.
[29] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D638–D646, Jan. 2023.
[30] H. Zou, "Pan-cancer assessment of mutational landscape in intrinsically disordered hotspots reveals potential driver genes," *Nucleic Acids Res.*, vol. 50, no. 9, p. e49, 2022.
[31] J. Liu, F. Ma, Y. Zhu, N. Zhang, L. Kong, J. Mi, H. Cong, R. Gao, M. Wang, and Y. Zhang, "MaxCLK: Discovery of cancer driver genes via maximal clique and information entropy of modules," *Bioinformatics*, vol. 39, no. 12, p. 737, Dec. 2023.
[32] W. Peng, Q. Tang, W. Dai, and T. Chen, "Improving cancer driver gene identification using multi-task learning on graph convolutional network," *Briefings Bioinf.*, vol. 23, no. 1, p. 432, Jan. 2022.

YANJIE REN received the M.S. degree in telecommunications engineering from Beijing University of Posts and Telecommunications, in 2012. He is currently pursuing the Ph.D. degree in computer science with Universiti Teknologi Malaysia. He is currently a Lecturer with Hebei Institute of Mechanical and Electrical Technology. His main research interests include artificial intelligence, data mining, and bioinformatics.

AZLAN MOHD ZAIN (Member, IEEE) received the Ph.D. degree in computer science from Universiti Teknologi Malaysia, in 2010. He is currently a Professor in computer science with Universiti Teknologi Malaysia. His main research interests include artificial intelligence, modeling and optimization, machining, and statistical process control.

YAN ZHANG received the M.S. degree in computer engineering from Yanshan University, in 2011. He is currently a Lecturer in computer science with Hebei Institute of Mechanical and Electrical Technology. His main research interests include artificial intelligence and data mining.

ROZITA ABDUL JALIL received the master's degree in computer science from Universiti Teknologi Malaysia, in 2005, and the Ph.D. degree in educational technology, in 2022. She is currently a Senior Lecturer with Universiti Tun Hussein Onn Malaysia (UTHM). Her research interests include machine learning, data science, educational technology, cognitive science, and instructional design.

MAHADI BAHARI received the Ph.D. degree from Brunel University London, U.K. He is currently an Associate Professor in information systems (IS). He has a background in IS and computing. He is also the Deputy Dean (Research, Innovation, and Development) with the Faculty of Management, Universiti Teknologi Malaysia. His research over the past decade has focused on the adoption and implementation of emerging and cutting-edge technologies, including artificial intelligence. More recently, his research interests include shifted toward marketing intelligence, controlling social media addiction, and readiness for healthcare IS.

NORFADZLAN BIN YUSUP received the Doctor of Philosophy degree in computer science from Universiti Teknologi Malaysia (UTM), in 2022. He is currently a Senior Lecturer with the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS). His research interests include machining and optimization, nature-inspired algorithms, machine learning, system development, and intelligent systems.

MAZLINA ABDUL MAJID received the Ph.D. degree in computer science from the University of Nottingham, U.K., in 2011. She is currently a Professor in computer science with Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), Malaysia. Her research interests include artificial intelligence, simulation modeling and optimization, data analytics, and software usability.

AZURAH A. SAMAH received the M.Sc. degree in operational research from Southampton U.K., in 1996, and the Ph.D. degree in operations management from Salford University, U.K., in 2010. She is currently an Associate Professor with the Faculty of Computing, Universiti Teknologi Malaysia, Malaysia. Formerly, she was the Head of the Artificial Intelligence and BioInformatics Research Group (AIBIG). Her research interests include simulation modeling, artificial intelligence, machine learning, deep learning, image processing, and operations management.

DIDIK DWI PRASETYA (Associate Member, IEEE) received the Ph.D. degree in information engineering from Hiroshima University, Japan. He is currently an Associate Professor with the Department of Electrical Engineering and Informatics, State University of Malang, Indonesia. His research interests include blossom in the fertile fields of knowledge engineering and data science, educational data science, data mining, learning analytics, and artificial intelligence (AI). Committed to research productivity and the publication of valuable papers, he increasingly demonstrates his academic involvement through his active role as a reviewer for reputable international journals.

NURHAFIZAH MOZIYANA MOHD YUSOP received the B.Sc. degree in computer science from Universiti Teknologi Malaysia, the master's degree in information technology (industrial computing), and the Ph.D. degree in industrial computing from Universiti Kebangsaan Malaysia. She is currently an Associate Professor with the Faculty of Defence Science and Technology, National Defence University of Malaysia (NDUM). Her research interests include numerical, artificial intelligence, assistive technology, and data analytics.