Specific Aims

The proposed research initiative aims to deepen our understanding of NSCLC progression in Black Americans, a demographic often underrepresented in cancer research. By implementing and critically evaluating two distinct computational frameworks — the established methods from Gerstung *et al.*, and the new GRITIC tool as described in the Baker *et al.*, and a novel approach integrating a HMM with machine learning — the proposal sets out to develop a comprehensive method to determine mutation timing and evolutionary patterns in NSCLC.

Aim 1: Apply existing timing methods to a new dataset of Black Americans with NSCLC to establish a baseline *Hypothesis: Applying the Gerstung et al. method and the GRITIC method to a new dataset will provide foundational insights into the mutation and evolutionary dynamics of NSCLC in Black Americans.*

- Acquire and preprocess the data for analysis
- Implement the Gerstung *et al.* and GRITIC methods to analyze the mutational and evolutionary landscape specific to Black Americans with NSCLC
- Identify key mutation timings and evolutionary patterns, and document these as the baseline for subsequent comparison using both methods

Aim 2: Develop and apply an advanced HMM integrated with machine learning to the same dataset to enhance mutation timing accuracy and uncover complex patterns *Hypothesis: An integrated HMM-machine learning model will reveal more detailed and accurate patterns of mutation timing and evolution in NSCLC, surpassing traditional and newly applied methods.*

- Design a hybrid framework that utilizes LSTM for feature extraction that feeds into an HMM to enhance the accuracy of state predictions and mutation timing
- Apply the new model to the dataset, focusing on improving the resolution of mutation timings and revealing hidden evolutionary pathways
- Compare the outcomes with those obtained from the Gerstung *et al.* method and the GRITIC method to evaluate improvements and gain insights into model efficacy

Aim 3: Benchmark the integrated HMM and machine learning model and extend the application to other types of cancer for comparative studies and generalizability assessment *Hypothesis: Benchmarking the integrated HMM and machine learning model against established methods including the GRITIC method will demonstrate its broad applicability and provide insights into the general mechanisms of cancer evolution, while establishing its comparative advantage over existing models.*

- Benchmark the new model against established methods including the GRITIC and Gerstung *et al.* methods using the PCAWG and NSCLC datasets to document performance metrics such as accuracy, computational efficiency, and adaptability
- Apply the benchmarked model to additional cancer datasets, focusing on those with significant disparities in diagnosis, treatment, or outcomes, to identify unique and shared mutation timings and evolutionary patterns
- Analyze and compare the results across different types of cancer to evaluate the model's effectiveness and adaptability in various oncological contexts

This project aims to revolutionize our understanding of NSCLC progression through cutting-edge advancements in modeling cancer evolution. By fulfilling these aims, the initiative is poised to bridge critical knowledge gaps in oncology, providing new analytical tools that map the evolution of cancer with unmatched precision. These advancements are expected to enhance the timing and specificity of treatment interventions and to lay the groundwork for tailored therapeutic strategies that are finely adjusted to the genetic and environmental profiles of individual patients, with special attention to the unique needs of Black communities. Ultimately, this research will catalyze significant improvements in early cancer detection, thereby transforming patient care and outcomes.

Research Strategy

Significance

Lung Cancer and Environmental Exposures Lung cancer remains the leading cause of cancer-related mortality worldwide, with non-small cell lung cancer (NSCLC) accounting for 85% of lung cancer cases in the US.[1] Akin to tobacco smoking, exposure to the complex mixture of air pollution, particularly fine particulate matter ($PM_{2.5}$) and nitric oxide (NO), poses a major risk factor for developing lung cancer. In heavily polluted cities like Los Angeles, exposure to these pollutants significantly increases the risk of developing lung cancer.[2,3] In 2014, the Nurse's Health Study found that living within 200 meters of a highway and a $10\ \mu g/m^3$ increase in $PM_{2.5}$ levels were associated with an increased risk of lung cancer (HR = 1.57; 95% CI: 1.26, 1.77).[4] Furthermore, a 2019 meta-analysis estimated that a $10\ \mu g/m^3$ increase in $PM_{2.5}$ exposure in Europe and North America increased lung cancer risk by 25%.[5]

Despite the clear evidence linking air pollution exposure to elevated lung cancer risk, the precise molecular mechanisms by which these complex pollutant mixtures initiate and promote NSCLC remain poorly understood, representing a critical knowledge gap. This study will investigate lung cancer in African Americans/Blacks (Black Americans), an understudied group that exhibits a high prevalence of aggressive, early-onset tumors that are often driven by distinct molecular profiles like EGFR mutations.[6] Elucidating the environmental drivers and biological pathways of lung carcinogenesis in this subgroup could reveal novel diagnostic approaches.

Addressing Lung Cancer Inequities in Black Americans Although Black Americans have lower smoking rates compared to non-Hispanic Whites, they experience significantly higher lung cancer incidence and mortality rates, especially among men.[7–9] This disparity is striking, as Black Americans tend to initiate smoking later in life and consume fewer cigarettes compared to their White counterparts.[8,10] Black women, despite smoking fewer cigarettes, have the same or higher incidence of lung cancer as White women.

Current lung cancer screening guidelines based on pack-years and age[11] fail to adequately identify Black Americans at risk. Black Americans are diagnosed with lung cancer at a significantly younger age than Whites, often before reaching the screening threshold of 30 pack-years or age 55.[12] The molecular drivers underlying these aggressive, early-onset lung cancers in the Black population remain unclear. However, disparities in environmental exposures, particularly air pollution, are suspected to play a role.[6] Evidence shows that Black Americans are consistently exposed to significantly higher levels of $PM_{2.5}$ and NO compared to non-Hispanic Whites. This study will utilize a multi-regional cohort of non-smokers, former smokers, and smokers, to identify the molecular connections between air pollutants and lung cancer in Black Americans.

Furthermore, existing studies do not account for how social determinants of health in Black Americans may modulate susceptibility to cancers.[9] Addressing this gap is crucial for accurately assessing risk and developing prevention strategies in diverse populations.

Characterization of Environmental Exposure Outdoor air pollution, including $PM_{2.5}$, is classified as a Group 1 carcinogen by the International Agency for Research on Cancer (IARC).[13] Past studies demonstrate a clear link between residing near major roadways and an elevated risk of developing lung cancer.[13] Exhaust from combustion engines releases a mixture of carcinogenic compounds into the atmosphere near major roadways. These pollutants include polycyclic aromatic hydrocarbons (PAHs), nitrogen oxides, and toxic heavy metals such as arsenic, nickel, and lead.[14] Previous studies have attempted to map air pollution levels using census tract data. However, these methods only detect a limited subset of pollutants, failing to capture the full complexity of environmental pollutants. Moreover, existing research does not account for how rising global temperatures associated with climate change may alter the chemical composition and carcinogenic potency of air pollution over time. Another major shortcoming is the lack of integration of social determinants of health, such as obesity, diabetes, and chronic inflammatory conditions, which may exacerbate susceptibility to cancer.

Harnessing Advanced Computational Models for Precise Mutation Timing and Cancer Progression Analysis Gerstung *et al.*, applies a suite of sophisticated computational tools including cancerTiming, MutationTimeR, PhylogicNDT SinglePatientTiming, and PhylogicNDT LeagueModel, to analyze cancer progression and mutation timing. Another recent advancement in genomic analysis is the Gain Route Identification and Timing In Cancer (GRITIC) method, detailed in Baker *et al.* This method employs advanced computational techniques to analyze complex genetic variations and is particularly adept at handling large-scale genomic datasets. The strength of GRITIC lies in its ability to time sequential copy number gains with high accuracy. However, the sophisticated Markov Chain Monte Carlo (MCMC) approach becomes computationally intractable for large copy numbers states (≥ 9).[16]

These methods utilize advanced probabilistic and statistical techniques to analyze mutational landscapes and identify evolutionary patterns in cancer genomes. However, these complex Bayesian inferences and MCMC approaches can be computationally intensive, limiting their use in large-scale or real-time scenarios. In addition, tools such as CancerTiming, MutationTimeR, and GRITIC focus on estimating the timing of clonal chromosomal gains and mutations but often require assumptions about mutation rates and copy number states that may not hold in all scenarios.

Hidden Markov Models (HMMs) are useful for modeling time-series data where the states of the system are hidden and must be inferred through observable events. HMMs capture transitions between hidden states based on observed data,

making them highly effective for sequential prediction tasks such as understanding disease progression.

Long Short-Term Memory (LSTM) networks are a specialized type of neural network that are highly effective in processing and retaining information across lengthy data sequences. This feature is particularly well-suited for modeling the complex evolutionary trajectories that occur in cancer. This allows the LSTM to dynamically learn from evolving genetic sequences and adapt to new patterns as they emerge.

Combining an LSTM with an HMM will enable dynamic modeling capabilities from LSTMs with the probabilistic modeling strengths of HMMs, with the potential to significantly improve cancer evolution studies. LSTMs provide a deeper and more nuanced understanding of long-term dependencies and non-linear interactions in genomic sequences, offering transformative improvements in the timing and ordering of mutations during cancer progression. When integrated with HMMs, which effectively model hidden states and transition probabilities, this approach allows for a more dynamic analysis that can adapt to new data, enhance prediction accuracy, and uncover hidden evolutionary pathways in cancer progression. This makes the LSTM-HMM model particularly powerful for predicting disease trajectories and improving treatment strategies, offering a superior alternative to traditional and current computational methods used in cancer genomics.

Potential for Transformative Impact This study will employ advanced geospatial methods to quantify individual exposures to air pollutants in Black communities in LA, Chicago, New Orleans, Charlestown SC, Richmond VA, and Rochester NY. Crucially, it will integrate this environmental exposure data with social determinants of health and biological factors that modulate disease susceptibility in these communities. Black populations in LA have historically faced disproportionately higher exposure to air pollution due to factors like redlining, the placing of industrial facilities near their neighborhoods, and a lack of green spaces. Despite having some of the lowest rates of smoking in the US, LA suffers from some of the worst highway-generated air pollution. By precisely characterizing these elevated exposures and combining them with data on obesity, diabetes, chronic inflammation, and other risk factors prevalent in Black communities, the goal is to develop a comprehensive analysis that elucidates how environmental drivers interact synergistically with social and biological parameters to initiate and promote aggressive, early-onset NSCLC in this population.

Integrating these assessments with the advanced capabilities of a hybrid HMM-LSTM model, the study will enhance the precision in mapping the timing of mutational events and uncover complex patterns of NSCLC progression. This multidisciplinary approach, which combines external exposure assessments with internal susceptibility factors, is poised to provide novel mechanistic insights into the environmental carcinogenesis pathways that contribute to the excessive lung cancer burden observed in Black communities. By correlating precise air pollution exposure data with epidemiological cohorts and molecular tumor profiling from Black NSCLC patients, the research will forge a comprehensive model of how environmental toxins catalyze lung carcinogenesis amidst the backdrop of social and biological vulnerabilities in this underserved population. We anticipate that this innovative approach will provide new insights into the role of air pollution in the development of NSCLC among Black Americans. This will help us develop targeted prevention, early detection, and treatment strategies. This is increasingly vital as, despite overall declining lung cancer rates, the incidence of NSCLC among women of color is rising in LA and similar urban areas.

Innovation

This proposal introduces several innovative elements to redefine our understanding of the role of air pollution in lung cancer in Black Americans. Traditional studies have often relied on broad regional data like census tract pollution maps that do not adequately capture individual exposures or the complex nature of environmental pollutants. In contrast, this study utilizes advanced geospatial monitoring to quantify personal exposures to $PM_{2.5}$, PAHs, NO, and metals, which are crucial for establishing clear exposure-response relationships.

A novel aspect of this proposal is the integration of an LSTM layer with an HMM. This hybrid approach is expected to enhance the precision of timing mutational events in NSCLC. The LSTM component is adept at analyzing long sequences of data, capturing temporal dependencies and patterns that may be indicative of mutational triggers linked to environmental factors. This capability is combined with the probabilistic power of HMMs, which model the hidden states and transition probabilities, offering a detailed temporal analysis of mutational sequences. Together, they provide a comprehensive view of how mutations develop and progress in response to specific environmental exposures.

Under the mentorship of Dr. Paul Spellman, a leader in identifying mutational signatures linked to specific carcinogens, this research will employ cutting-edge genomic techniques to meticulously chart the molecular alterations induced by environmental exposures. The LSTM-HMM allows for a dynamic and nuanced analysis, setting this approach apart by providing a robust framework for directly connecting specific components of air pollution to the pathogenesis of lung cancer in Black Americans. The hybrid model will rigorously define the temporal and causal relationships between environmental exposures and mutational events, establishing a direct and unambiguous link between air pollution and lung cancer development.

Overall, this proposal leverages sophisticated exposure monitoring integrated with advanced computational modeling to better understand environmental lung carcinogenesis. This multi-disciplinary approach, combining high-resolution environmental data with innovative computational techniques, has the potential to transform our understanding of the role of air pollution in NSCLC and identify new avenues for prevention and early detection in Black Americans.

Approach

Aim 1: Apply existing timing methods to a new dataset of Black Americans with NSCLC to establish a baseline

Rationale Aim 1 focuses on applying the tools used in Gerstung *et al.* alongside the GRITIC method to a newly compiled dataset of Black Americans diagnosed with NSCLC. This approach is motivated by the need to establish a reliable baseline of mutation timing and evolutionary patterns that are specific to this demographic, which is often underrepresented in cancer research. The Gerstung *et al.* method provides a proven framework for analyzing mutational signatures and evolutionary trajectories, providing insights into the genetic dynamics of cancer progression. The GRITIC method complements this analysis by providing precise timings of sequential copy number gains, via a sophisticated MCMC algorithm to unravel complex genetic variations. Together, these methods will allow us to map out a mutational landscape, providing a foundational analysis, that will be used to compare methodological efficacy and validate the model developed in Aim 2.

1.1. Acquire and preprocess the data for analysis This stage involves collaborating with regional medical centers and cancer research networks to collect extensive genomic data from NSCLC patients, particularly those of Black American descent. The data will undergo rigorous cleaning and processes to remove biases and errors. This includes alignment to reference genomes, removing duplicate reads, correcting sequencing errors, and accounting for mutations that occur in healthy cells. We will also implement stringent quality control measures to validate the integrity and completeness of the data, ensuring that it meets the analytical requirements of the computational methods to be applied. While it may seem trivial, this step is expected to take the longest, as it is not always immidiately clear what preprocessing steps are required.

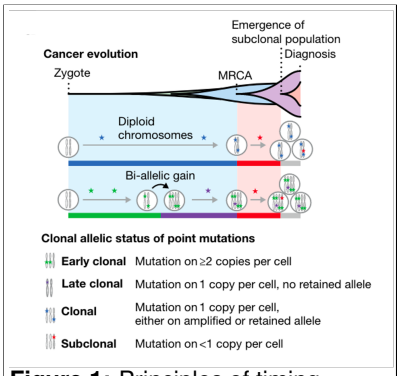

Figure 1: Principles of timing mutations and copy number gains based on whole-genome sequencing. Figure from Gerstung *et al.*

1.2. Implement the Gerstung *et al.* and GRITIC methods to analyze the mutational and evolutionary landscape specific to Black Americans with NSCLC In this phase, we will apply the Gerstung *et al.* method to assess mutational signatures and evolutionary trajectories. This method will help us statistically model mutation rates and clonal evolution patterns to infer the historical development of tumors. We will also use the GRITIC method to analyze sequential copy number gains. GRITIC implements an advanced MCMC algorithm to infer the timing and progression of these genetic changes. The integration of findings from both methods will allow us to develop a baseline anaysis of tumor evolution in the NCSLC dataset, focusing on identifying critical mutational events and their timings.

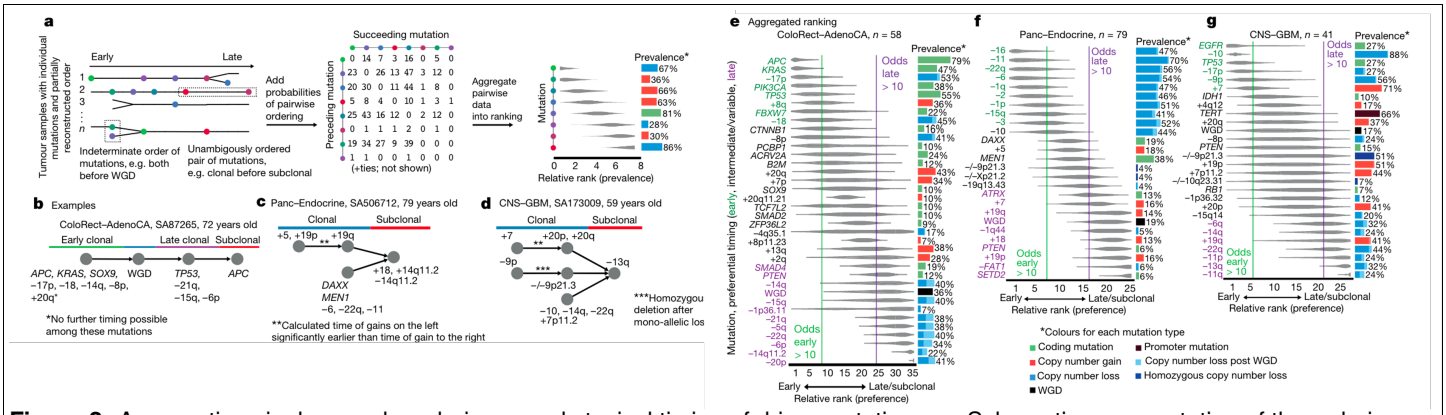

Figure 2: Aggregating single-sample ordering reveals typical timing of driver mutations. a, Schematic representation of the ordering process. b–d, Examples of individual patient trajectories (partial ordering relationships), the constituent data for the ordering model process. e–g, Preferential ordering diagrams for colorectal adenocarcinoma (ColoRect–AdenoCA) (e), pancreatic neuroendocrine cancer (Panc–Endocrine) (f) and glioblastoma (CNS–GBM) (g). Probability distributions show the uncertainty of timing for specific events in the cohort. Figure from Gerstung *et al.*

1.3. Identify key mutation timings and evolutionary patterns, and document these as the baseline for subsequent comparison using both methods After conducting the computational analyses, we will synthesize the results to highlight key mutational timings and evolutionary patterns. A comparative analysis will be performed to align our findings with existing literature on NSCLC, especially focusing on unique trends and discrepancies relevant to the Black American population. Detailed reports and visual representations, such as evolutionary trees and mutation timelines, will be prepared to document these baseline findings comprehensively. This documentation will serve as a reference for subsequent analyses.

Challenges & Alternative Approaches We anticipate challenges such as high inter-individual variability in mutation rates and evolutionary patterns. To address this, we plan to incorporate a larger sample size and employ robust statistical methods to ensure the generalizability of our findings. The computational demands of the GRITIC method will be managed by leveraging high-performance computing resources and optimizing algorithm parameters to enhance computational efficiency without compromising accuracy. We will also ensure seamless integration of results from different computational methods by using standardized data formats and developing custom scripts for data merging and analysis.

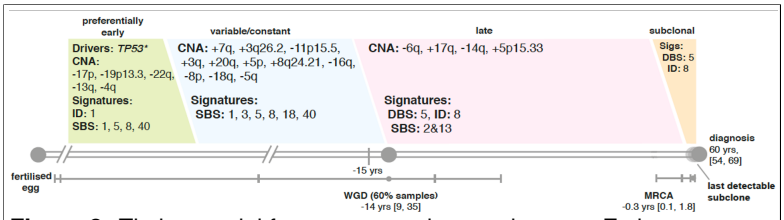

Figure 3: Timing model for serous ovarian carcinomas. Estimates on when major events occur during tumor development are provided.[15]

Aim 2: Develop and apply an advanced HMM integrated with machine learning to the same dataset to enhance mutation timing accuracy and uncover complex patterns

Rationale In the realm of cancer research, particularly in studies like those conducted by Gerstung *et al.*, traditional methods for modeling mutation timings often rely on statistical models. These models typically encompass a variety of approaches ranging from regression-based frameworks to simpler probabilistic models like Markov chains. These models are well-established for their interpretability and methodological transparency but may lack the nuanced capacity to handle complex, high-dimensional, and non-linear patterns that characterize genetic data and cancer progression.

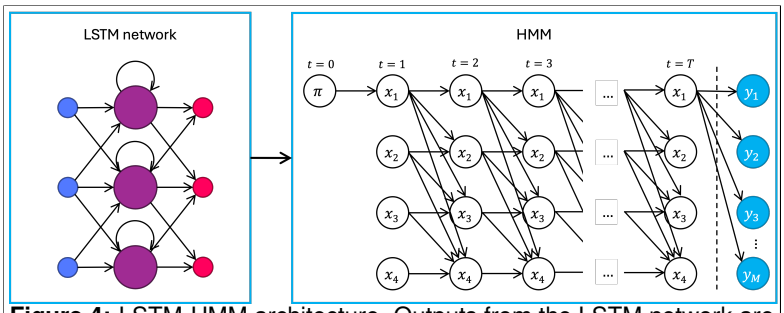

Figure 4: LSTM-HMM architecture. Outputs from the LSTM network are used to inititate the HMM.

Unlike conventional statistical models that may not capture complex dependencies, the LSTM component of the hybrid model can analyze sequential and temporal data over extended periods. This capability allows it to discern underlying patterns that are predictive of disease progression that might be overlooked by other methods. By integrating LSTM-derived insights with an HMM, the hybrid model enriches the probabilistic modeling of hidden states. The LSTM layer enables the interpretation of details and complexities within sequential data that may not be immediately apparent or directly measurable. Using the LSTM outputs in the HMM model improves the HMM's ability to predict state transitions and better capture the complex interactions in the mutation processes.

Utilizing an LSTM enables the model to learn from and adapt to new data dynamically, a significant advantage over traditional models that may require static reconfiguration or retraining. The integration of LSTM with HMM not only enhances accuracy but also provides finer resolution in the timing of mutations, offering detailed insights that are crucial for effective treatment planning and personalized medicine.

2.1. Design a hybrid framework that utilizes LSTM for feature extraction that feeds into an HMM to enhance the accuracy of state predictions and mutation timing To harness the predictive power of advanced machine learning for understanding disease progression, we propose developing an LSTM network specifically designed to analyze biological sequences. The LSTM layer will extract temporal features and patterns that indicate shifts in disease progression. The LSTM will output detailed state prediction probabilities and feature vectors that encapsulate the dynamic characteristics of the disease, providing a rich dataset that reflects both the current state and the likely future states of the disease. These outputs from the LSTM will then be fed into the HMM. Specifically, the state prediction probabilities generated by the LSTM will be used as emission probabilities in the HMM. This step is expected to refine the HMM's ability to map observed data to the correct hidden states. Additionally, the feature vectors derived from the LSTM will serve as observations in the HMM. This integration enhances the HMM's capability to accurately model transitions between different disease states, leveraging the detailed context provided by the LSTM to improve both the accuracy and reliability of the disease progression model.

2.2. Apply the new model to the dataset, focusing on improving the resolution of mutation timings and revealing hidden evolutionary pathways In this aim, we will implement and optimize both LSTM and HMM models within a unified framework, to ensure robust predictions of mutation events. To achieve this, we will initially pre-train the LSTM on a designated subset of the PCAWG dataset. This preliminary step is designed to stabilize the LSTM's feature extraction capabilities before it is fully integrated with the HMM. After we train the LSTM, we will use the outputs of the LTSM to dynamically adjust the parameters of the HMM. This adjustment process will utilize a combined training approach that integrates backpropagation for optimizing the LSTM alongside the Baum-Welch algorithm for refining the HMM. To validate the model's robustness, we will perform cross-validation to prevent overfitting and to ensure that the model generalizes effectively across different subsets

of data. Lastly, we will continuously optimize the model parameters, focusing on improving key performance metrics such as accuracy, sensitivity, and specificity in predicting mutation timing. This thorough approach to training and validation aims to create a predictive model that is both precise and reliable in its application to real-world clinical data.

2.3. Compare the outcomes with those obtained from the Gerstung *et al.* method and the GRITIC method to evaluate improvements and gain insights into model efficacy Our objective is to apply the fully integrated LSTM-HMM model to the entire NSCLC dataset to assess improvements in the accuracy of mutation timing and to uncover hidden evolutionary patterns. By applying this hybrid model, we aim to focus on achieving detailed and high-resolution timing of mutation events, which is critical for advancing our understanding of cancer progression. Once applied, we will conduct a comparative analysis to juxtapose the mutation timings and evolutionary patterns derived from our hybrid model against those obtained using the traditional Gerstung et al. method. This comparison will not only highlight the differences and improvements brought about by our model but also allow us to evaluate the added value of integrating LSTM with HMM. Specifically, we will assess enhancements in resolution, adaptability, and predictive accuracy, thereby demonstrating the hybrid model's potential to significantly advance the field of cancer genomics.

Challenges & Alternative Approaches In our approach to model complexity and data sparsity, it's important to note that HMMs are inherently proficient at handling sparse data, providing a robust fallback mechanism when large datasets are not available. This capability allows us to potentially minimize the reliance on extensive machine learning processes in scenarios where data is limited, while still maintaining the flexibility to fully leverage more complex machine learning techniques, such as those involving LSTM networks, as richer datasets become available. By designing the system with this dual capability, we ensure that the model remains effective under varied data conditions, offering both simplicity and adaptability. This approach not only safeguards against overfitting but also ensures that the model can dynamically scale its complexity based on the volume and detail of the data it processes.

Aim 3: Benchmark the integrated HMM and machine learning model and extend the application to other types of cancer for comparative studies and generalizability assessment

Rationale Aim 3 of our project involves the crucial step of benchmarking the LSTM-HMM hybrid model against established computational methods such as those from Gerstung et al. and the GRITIC method. Furthermore, we plan to extend the application of this model to other types of cancer to assess its generalizability and efficacy across various oncological contexts. This approach is designed to solidify the model's robustness and adaptability and to potentially set a new standard in the precision modeling of cancer evolution.

3.1. Benchmark the new model against established methods including the GRITIC and Gerstung *et al.* methods using the PCAWG and NSCLC datasets to document performance metrics such as accuracy, computational efficiency, and adaptability We will start by implementing the LSTM-HMM model on the NSCLC dataset that has been analyzed in Aim 1 and Aim 2. The primary goal here is to directly compare the performance of our model against the outcomes derived from the Gerstung et al. and GRITIC methods. Performance metrics such as accuracy, computational efficiency, and adaptability will be meticulously documented. We will use statistical measures like the area under the ROC curve (AUC), confusion matrices for classification accuracy, and time-efficiency analyses to quantitatively assess each model's performance.

3.2. Apply the benchmarked model to additional cancer datasets, focusing on those with significant disparities in diagnosis, treatment, or outcomes, to identify unique and shared mutation timings and evolutionary patterns After benchmarking, we will apply the LSTM-HMM model to additional cancer datasets. These datasets will be selected based on their variance in diagnosis, treatment outcomes, and genetic diversity to ensure a comprehensive test of the model's applicability. For each new type of cancer analyzed, we will adjust and fine-tune the model parameters to accommodate different genetic signatures and mutation rates. This phase will help in identifying unique and shared mutation timings and evolutionary patterns across cancers, which could be crucial for understanding broad and specific pathways of oncogenesis.

3.3. Analyze and compare the results across different types of cancer to evaluate the model's effectiveness and adaptability in various oncological contexts Each application of the model will be followed by a detailed comparative analysis where results from the LSTM-HMM model will be juxtaposed against those obtained using traditional methods. This will allow us to evaluate the added value of our approach in terms of enhanced resolution, predictive accuracy, and the ability to adapt to different types of cancer data. Insights gained from these comparisons will be critical in demonstrating the model's effectiveness and could pave the way for its adoption in clinical settings.

Challenges & Alternative Approaches One anticipated challenge is the variance in data quality and completeness across different cancer datasets, which could affect model performance. To mitigate this, we will employ data augmentation techniques and sophisticated imputation methods to ensure data integrity. Additionally, the computational demands of applying LSTM-HMM to diverse and extensive datasets will be managed through the use of scalable cloud computing resources and optimizing the model's architecture for high-performance computing environments.

Glossary

References

[1] Julian R. Molina, Ping Yang, Stephen D. Cassivi, Steven E. Schild, and Alex A. Adjei. “Non–Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship”. In: *Mayo Clinic proceedings. Mayo Clinic* 83.5 (May 2008). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2718421/, pp. 584–594. ISSN: 0025-6196.

[2] Christine D. Berg, Joan H. Schiller, Paolo Boffetta, Jing Cai, Casey Connolly, Anna Kerpel-Fronius, Andrea Borondy Kitts, David C.L. Lam, Anant Mohan, Renelle Myers, Tejas Suri, Martin C. Tammemagi, Dawei Yang, and Stephen Lam. “Air Pollution and Lung Cancer: A Review by International Association for the Study of Lung Cancer Early Detection and Screening Committee”. In: *Journal of Thoracic Oncology* 18.10 (Oct. 2023), pp. 1277–1289. ISSN: 15560864. DOI: 10.1016/j.jtho.2023.05.024.

[3] Yanqian Huang, Meng Zhu, Mengmeng Ji, Jingyi Fan, Junxing Xie, Xiaoxia Wei, Xiangxiang Jiang, Jing Xu, Liang Chen, Rong Yin, Yuzhuo Wang, Juncheng Dai, Guangfu Jin, Lin Xu, Zhibin Hu, Hongxia Ma, and Hongbing Shen. “Air Pollution, Genetic Factors, and the Risk of Lung Cancer: A Prospective Study in the UK Biobank”. In: *American Journal of Respiratory and Critical Care Medicine* 204.7 (Oct. 2021), pp. 817–825. ISSN: 1073-449X. DOI: 10.1164/rccm.202011-4063OC.

[4] Robin C. Puett, Jaime E. Hart, Jeff D. Yanosky, Donna Spiegelman, Molin Wang, Jared A. Fisher, Biling Hong, and Francine Laden. “Particulate Matter Air Pollution Exposure, Distance to Road, and Incident Lung Cancer in the Nurses’ Health Study Cohort”. In: *Environmental Health Perspectives* 122.9 (Sept. 2014), pp. 926–932. ISSN: 1552-9924. DOI: 10.1289/ehp.1307490.

[5] Hung-Ling Huang, Yung-Hsin Chuang, Tzu-Hsuan Lin, Changqing Lin, Yen-Hsu Chen, Jen-Yu Hung, and Ta-Chien Chan. “Ambient Cumulative PM2.5 Exposure and the Risk of Lung Cancer Incidence and Mortality: A Retrospective Cohort Study”. In: *International Journal of Environmental Research and Public Health* 18.23 (Nov. 2021), p. 12400. ISSN: 1661-7827. DOI: 10.3390/ijerph182312400.

[6] Anita Marcinkiewicz, Aleksandra Ochotnicka, Karolina Borowska-Waniak, Kinga Skorupińska, Dominik Michalik, and Maja Borowska. “THE IMPACT OF AIR POLLUTION ON THE OCCURRENCE OF LUNG CANCER: A LITERATURE REVIEW”. In: *Archiv Euromedica* 13.5 (Oct. 2023). ISSN: 2199885X. DOI: 10.35630/2023/13/5.507.

[7] Nadine Belony, Bing Ren, Phuc Pham, Matthew Gregory, Pablo E. Puente, Nazarius S. Lamango, Ite A. Offringa, and Yong Huang. “Abstract 4038: Study of Therapeutic Effects of Polyisoprenylated Cysteinyl Amide Inhibitors on Lung Cancer Cells of Black Patients Using 3D-Printed Alveolar Model”. In: *Cancer Research* 82.12_Supplement (June 2022), p. 4038. ISSN: 0008-5472. DOI: 10.1158/1538-7445.AM2022-4038.

[8] Jose Thomas Thaiparambil, Zheng Yin, and Randa El-Zein. “Abstract 1948: Novel Insights into Lung Cancer & Chronic Obstructive Pulmonary Disease Racial Disparities”. In: *Cancer Research* 83.7_Supplement (Apr. 2023), p. 1948. ISSN: 0008-5472. DOI: 10.1158/1538-7445.AM2023-1948.

[9] Celina I. Valencia, Francine C. Gachupin, Yamilé Molina, and Ken Batai. “Interrogating Patterns of Cancer Disparities by Expanding the Social Determinants of Health Framework to Include Biological Pathways of Social Experiences”. In: *International Journal of Environmental Research and Public Health* 19.4 (Feb. 2022), p. 2455. ISSN: 1661-7827. DOI: 10.3390/ijerph19042455.

[10] Karyn Hede. “Drilling Down to the Causes of Racial Disparities in Lung Cancer”. In: *JNCI: Journal of the National Cancer Institute* 102.18 (Sept. 2010), pp. 1385–1387. ISSN: 0027-8874. DOI: 10.1093/jnci/djq371.

[11] Rebecca Landy, Li C. Cheung, Corey D. Young, Anil K. Chaturvedi, and Hormuzd A. Katki. “Absolute Lung Cancer Risk Increases among Individuals with $>$15 Quit-Years: Analyses to Inform the Update of the American Cancer Society Lung Cancer Screening Guidelines”. In: *Cancer* 130.2 (Jan. 2024), pp. 201–215. ISSN: 0008-543X. DOI: 10.1002/cncr.34758.

[12] Rafael Meza, Jihyoun Jeon, Iakovos Toumazis, Kevin Ten Haaf, Pianpian Cao, Mehrad Bastani, Summer S. Han, Erik F. Blom, Daniel E. Jonas, Eric J. Feuer, Sylvia K. Plevritis, Harry J. de Koning, and Chung Yin Kong. “Evaluation of the Benefits and Harms of Lung Cancer Screening With Low-Dose Computed Tomography: Modeling Study for the US Preventive Services Task Force”. In: *JAMA* 325.10 (Mar. 2021), pp. 988–997. ISSN: 1538-3598. DOI: 10.1001/jama.2021.1077.

[13] Shilpa N. Gowda, Anneclaire J. DeRoos, Rebecca P. Hunt, Amanda J. Gassett, Maria C. Mirabelli, Chloe E. Bird, Helene G. Margolis, Dorothy Lane, Matthew R. Bonner, Garnet Anderson, Eric A. Whitsel, Joel D. Kaufman, and Parveen Bhatti. “Ambient Air Pollution and Lung Cancer Risk among Never-Smokers in the Women’s Health Initiative”. In: *Environmental Epidemiology* 3.6 (Oct. 2019), e076. ISSN: 2474-7882. DOI: 10.1097/EE9.0000000000000076.

[14] Xian-Jun Yu, Min-Jun Yang, Bo Zhou, Gui-Zhen Wang, Yun-Chao Huang, Li-Chuan Wu, Xin Cheng, Zhe-Sheng Wen, Jin-Yan Huang, Yun-Dong Zhang, Xiao-Hong Gao, Gao-Feng Li, Shui-Wang He, Zhao-Hui Gu, Liang Ma, Chun-Ming Pan, Ping Wang, Hao-Bin Chen, Zhi-Peng Hong, Xiao-Lu Wang, Wen-Jing Mao, Xiao-Long Jin, Hui Kang, Shu-Ting Chen, Yong-Qiang Zhu, Wen-Yi Gu, Zi Liu, Hui Dong, Lin-Wei Tian, Sai-Juan Chen, Yi Cao, Sheng-Yue Wang, and Guang-Biao Zhou. “Characterization of Somatic Mutations in Air Pollution-Related Lung Cancer”. In: *EBioMedicine* 2.6 (June 2015), pp. 583–590. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2015.04.003.

[15] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G. Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C. Boutros, David D. Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D. Morris, Paul T. Spellman, David C. Wedge, and Peter Van Loo. “The Evolutionary History of 2,658 Cancers”. In: *Nature* 578.7793 (Feb. 2020), pp. 122–128. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1907-7.

[16] Toby M. Baker, Siqi Lai, Tom Lesluyes, Haixi Yan, Annelien Verfaillie, Stefan Dentro, Andrew R. Lynch, Amy L. Bowes, Nischalan Pillay, Adrienne M. Flanagan, Charles Swanton, Maxime Tarabichi, and Peter Van Loo. *The History of Chromosomal Instability in Genome Doubled Tumors*. Oct. 2023. DOI: 10.1101/2023.10.22.563273.