American Journal of Biomedicine and Pharmacy
https://biojournals.us/index.php/AJBP
ISSN: 2997-7177

Infrared Spectroscopy-Enabled Molecular Fingerprinting of Blood Samples Coupled with Machine Learning for Multiplex Disease Screening

Hawraa Abbas Alwan Aziz
Thi-Qar university college of pharmacy

Elham Adham Noaman Khidr
Knowledge University College of Science Department of Pathological Analysis

Azad Mahmoud Abdul Jabbar Haji
University of Mosul College of Science Department of Biology

Karar Saeed Kazim Muhammad
University of Babylon College of Science Department of Life Sciences, Microbiology Branch

Sondos Adnan Nouri Hazaa
University of Tikrit College of Science Department of Chemistry

Received: 2024, 15, Jun
Accepted: 2025, 21, Jul
Published: 2025, 26, Aug

Annotation: Molecular fingerprinting of blood samples, when effectively combined with advanced machine learning techniques, facilitates an innovative approach to label-free multiplex screening of a wide array of both acute and chronic diseases. Blood, as a readily accessible and abundant source of vital biochemical information, serves as an exceptionally ideal medium for implementing such comprehensive screening methods. The seamless integration of molecular fingerprinting and artificial intelligence capitalizes on intrinsic molecular signatures, thereby enabling the rapid assessment of disease presence. This supports extensive screening initiatives that are crucial in today's healthcare landscape. Moreover, this multifaceted approach transforms complex raw spectral data into clear and quantifiable molecular profiles. These profiles not only enhance the understanding of various

pathological conditions but also provide critical insights into diseases such as cardiovascular conditions, type-2 diabetes mellitus, a range of liver diseases, and different types of cancers. Importantly, this method effectively meets the pressing demand for scalable, cost-effective, and minimally invasive disease screening solutions, especially during resource-constrained scenarios, which may include pandemic outbreaks and public health emergencies. Overall, the fusion of molecular fingerprinting with machine learning holds immense potential to revolutionize disease detection and monitoring practices in the future.

1. Introduction

Existing diagnostics for infectious and inflammatory diseases typically rely on rapid antigen tests or molecular probing and recognition, which may not be available or affordable in many parts of the world [1]. Fundamental considerations for multiplex disease screening include rapid turnaround time, minimal sample preparation, label-free operation, robustness, cost-effectiveness, and ease of operation. The overarching aim is to provide a holistic view into the biochemical composition of a biological sample (or its so-called molecular fingerprint), wherein the disease-specific molecular profiles are often engraved. This capability opens an avenue for developing next-generation data-driven diagnostics to meet the future-required diagnostic challenges [2].

The present Report delineates a distinct methodology based upon the molecular fingerprinting of conveniently acquired blood samples, which, when augmented with a suitable machine learning pipeline, can be deployed towards multiplex screening and differentiation of highly different diseases (e.g., tuberculosis, pneumonia, sickle cell anaemia). The in vitro examination reveals that the 2D fingerprint also mirrors the propagation dynamics of the diseases in question. Furthermore, the in vivo investigation recapitulates some clinical observations in common inflammatory diseases. The collective raw data and source codes (Supplementary Section ARTICLE_CODE), along with the informed user protocols, will expedite the adoption of the technique by other researchers.

2. Background

Molecular diagnostics aims at detecting diseases using biological molecules such as DNA, RNA, proteins, and their interactions. Emerging technologies for rapid, high-throughput, and automated molecular fingerprinting of biological samples thus hold potential for point-of-care healthcare screening and multiplex diagnosis. Blood biomolecules are regarded as effective indicators for many diseases, and different blood-based molecular fingerprinting techniques have been proposed [1]. However, there is no generic workflow that incorporates blood-based molecular fingerprinting for multiplex disease screening and connects seamlessly to a machine learning (ML) framework. ML enables a system to learn and solve problems from past data and experiences without explicit programming. It has been employed in various fields ranging from natural language processing to computer vision, and even helps discover unknown patterns and hidden relationships in data to make scientific conclusions [3]. Knowledge extraction from large biological datasets such as molecular profiles, images, and clinical records may help to understand and analyze complex biological pathways and functions, and allow effective disease prediction and screening. Coupled with ML, biomolecular fingerprinting can harness unexploited

information in data and enable rapid, label-free phenotyping of biological samples. A few efforts have successfully demonstrated single-disease diagnosis using ML-assisted fingerprinting techniques. However, a generic procedure has yet to be established for multiplex screening—despite multifarious diseases afflicting humans, many exhibiting similar symptoms and comorbidities. The ambient information present within biological samples makes it challenging to decipher and build high-performance ML classifiers capable of differentiating multiple diseases simultaneously.

2.1. Overview of Molecular Fingerprinting

Molecular fingerprinting offers a rapid and efficient alternative to standard genotyping and phenotyping methods for the analysis of biological samples [1]. Genotyping and phenotyping — including proteomics, lipidomics and other "-omics" methods — are needed to close the gap between bench-to-bedside translations from basic science and clinical research. However, conventional methods remain relatively slow, complex or expensive for large-scale practical application. When coupled with fast, efficient multi-dimensional inverse Laplace decomposition (ILD), molecular fingerprinting enables the rapid extraction of multiple molecular information from a single one-dimensional measurement.

Applied to a small amount of blood sample, two-dimensional correlational spectroscopy generates a quiescent "molecular fingerprint" that identifies multiple coexisting constituents in a complex sample at low cost in minutes. When coupled with machine learning, the fingerprint can be classified swiftly without the need for sophisticated analyses. Potential applications at this early stage include molecular phenotyping of oxidation level and detection of haemoglobin variants, which are important for managing various clinical conditions and high-burden haemoglobinopathies.

Molecular fingerprinting can be implemented with a much smaller device footprint compared to conventional optical spectroscopy or mass spectrometry, providing a promising sensing technique for point-of-care-testing and screening of a wide range of complex biological fluids and materials. Two-dimensional nuclear magnetic resonance T1–T2 correlational spectroscopy constitutes one such platform, providing rapid, label-free and non-invasive molecular information decoding without requiring chemical reagents or additional laboratory preparations. The physical principle is based on the encoding of molecular dynamics on multi-nuclear-spin systems, which reveals intrinsic molecular fingerprints sensitive to composition and micro-structural environment.

2.2. Importance of Blood Samples in Disease Screening

The Food and Drug Administration (FDA) is charged with ensuring the safety of the over 15 million units of blood and blood components donated each year in the United States. Screening donated blood for infectious diseases transmissible through transfusion constitutes a critical safeguard. As a result, the United States currently possesses the safest blood supply in the world, with the FDA continuously revising standards governing the collection and processing of blood. Every donated unit undergoes tests that satisfy regulatory mandates for infectious diseases [4]. Modern clinical microbiology and virology have increasingly adopted molecular technologies, such as nucleic acid testing, to detect viral infections acquired so recently that conventional serologic methods cannot yet identify them. The implementation of molecular detection techniques enables a further significant reduction in the incidence of transfusion-transmitted disease.

2.3. Machine Learning in Healthcare

Machine learning is a subset of Artificial Intelligence (AI) that enables the development of enhanced algorithms that improve themselves through iteration using previous experience. With the exponential growth of healthcare data and physician demand, machine learning applications in healthcare have grown rapidly since 2015. Machine learning has been successfully applied to

both Image Processing and Biomedical Signal Classification tasks.

In imaging, deep learning has been deployed to detect features in chest x-rays, MRI imagery, CT scans, ultrasound, and pathology slides for COVID-19, cancer, Alzheimer's disease, among other conditions. The field of biomedical signal processing uses physiological signals such as ECG, EEG, and EMG. These signals reflect the electrical activity of the heart, brain, and muscles, respectively. Therefore, machine learning-based analysis of these signals can reveal how this electrical activity changes with disease progression.

3. Methodology

A standard operating procedure for rapid, label-free molecular phenotyping of blood by two-dimensional magnetic resonance (MR) relaxometry coupled with machine learning was established. The approach required 130 µl of raw sample, which was prepared, acquired, and analysed in less than 5 minutes. A high-throughput, low-cost MR setup conducted two-dimensional relaxation (T1–T2) measurements on whole blood, red blood cells, plasma, and saliva; extracted underlying sub-domains; and queried available databases for pattern recognition. The machine learning–molecular fingerprint method classified various biological states, including oxidative stress, inflammation, and ischaemia, with better accuracy than conventional supervised-learning techniques. Systematic comparison against standard biochemical assays demonstrated significantly improved capacities in screening, triaging, and continuous monitoring of multiple diseases, spanning the entire spectrum from oesophageal inflammation to cancer. [1] [3]

3.1. Sample Collection and Preparation

Ethics approval was obtained from the Institutional Review Boards of the National University of Singapore (reference code NUS-IRB-11-354). Whole blood and serum samples were collected from consenting volunteers. Fresh blood samples were collected into an ethylenediaminetetraacetic acid (EDTA) tube and stored at 4 °C; pre-existing blood samples stored in a heparinized tube at 4 °C were also utilized. Serum samples were stored at −20 °C prior to analysis. For sample preparation, 500 µL of whole blood was centrifuged for 10 min at 14,000×g. The plasma was removed, and the red blood cells (RBCs) at the bottom were washed twice with 500 µL of a phosphate-buffered saline (PBS) buffer (pH 7.4). After the final wash, the pellet was resuspended in 500 µL PBS and homogenized by gentle vortexing for NMR measurement.

Large-scale genetic epidemiological studies require high-quality analysis of samples such as blood or saliva from multiple patients, which is challenging at the point of care. To expand these studies' impact, minimal sample storage time and less complex extraction of DNA or RNA for downstream applications are necessary. Whole blood was collected from healthy volunteers and stored at 4 °C in ethylenediaminetetraacetic acid (EDTA) vacutainers until use for genomic DNA (gDNA) extraction.

A microfluidics-based system was developed for gDNA extraction from whole blood. A mixture of blood lysate, paramagnetic beads, and binding buffer was placed into an input well. gDNA-bound beads were pulled using a magnet through a wash buffer to an output well containing elution buffer, where DNA was eluted at 55 °C off the chip. The 40-minute protocol extracted gDNA from six samples simultaneously, requiring 4 µL of diluted blood and a total reagent volume of 75 µL per reaction. qPCR and spectrofluorimetry tested the purity and quantity of eluted gDNA. Bead transport and molecular diffusional analysis indicated an input of less than 4 ng of gDNA was optimal. No inhibitory transport affected qPCR, and samples were prepared for next-generation sequencing. The microfluidic extraction method is vital for future DNA-based point-of-care diagnostics and NGS workflows.

Fresh blood samples were collected into an EDTA tube and stored at 4 °C; pre-existing blood samples stored in a heparinized tube at 4 °C were also utilized. Serum samples were stored at

−20 °C prior to analysis. For sample preparation, 500 μL of whole blood was centrifuged for 10 min at 14,000×g. The plasma was removed, and the RBCs at the bottom were washed twice with 500 μL of PBS buffer (pH 7.4). After the final wash, the pellet was resuspended in 500 μL PBS and homogenized by gentle vortexing for NMR measurement [1].

3.2. Molecular Fingerprinting Techniques

Fingerprinting techniques are an essential step in capturing the molecular signatures of blood samples across a range of length scales. For DNA-level genotyping, mini-STR analysis offers dry chemistry in an evacuated tube. The STR profile is then compared with cell-line repositories for sample authentication. This approach remains an industry standard and can support multiplex assay development. For bulk phenotyping, a two-dimensional (2D) T1-T2 NMR correlational spectroscopy setup analyzes a single drop of blood [1]. The output of the spin-echo train is subjected to inverse Laplace decomposition for correlation of the resulting relaxation times, providing a rapid, label-free molecular fingerprint. At the protein level, a multiplexed proximity extension assay measures approximately 1500 proteins from a 5-μl sample. Antibody-oligonucleotide conjugates bind in pairs to protein targets; a qPCR readout quantifies the proximity probes and infers target protein abundance. Peptidoglycan from cell walls is detected by fluorescent staining and flow cytometry within the 6–40 kDa range of the serum proteome. An MLGA (multiplex ligation-dependent genome amplification) technique detects 60 indel polymorphisms with high tolerance for template degradation [3] ; this amplified multiplex amplicon mixture can be used for sample validation before re-sequencing. Cell and/or microparticle concentrations are also estimated by flow cytometry. These multiple approaches provide complementary fingerprints of molecular patterns at the various scales relevant to disease screening and diagnostics [5].

3.3. Machine Learning Algorithms Used

The study incorporates a variety of machine learning algorithms, selected for their capacity to analyze molecular fingerprints effectively and enable multiplex disease screening. Among these, support vector machines (SVM) stand out as a robust choice. SVMs employ kernel functions to project data into high-dimensional spaces, thereby facilitating linear separation of classes that may be non-linearly separable in the original feature space [6]. This method has demonstrated high classification accuracy across numerous biomedical tasks, including genetic disorder prediction and health-risk estimation [7]. Complementing SVMs, K-Nearest Neighbors (KNN) offers a straightforward, instance-based learning approach. By classifying samples based on proximity to labeled neighbors within the feature space, KNN maintains excellent performance in multiclass medical applications. Additional algorithms such as decision trees and logistic regression are deployed to provide further classification perspectives; decision trees facilitate rule-based interpretation of complex patterns, while logistic regression quantifies relationships between variables and binary outcomes. This suite of algorithms—emphasizing molecular fingerprint analysis rather than gene expression data—provides the analytical foundation for disease classification within the study.

4. Data Acquisition

The present discussion derives data from both laboratory recordings and published studies. The initial sample set comprises a human whole blood dataset of 218 individuals, encompassing healthy controls and patients diagnosed with various disease states across four categories: tuberculosis (n = 38), stroke (n = 54), meningitis (n = 35), and dengue (n = 25). The Institutional Review Board has granted approval for the human subject study, and all participants have provided written informed consent prior to enrollment [1]. Blood samples are processed within 5–10 minutes of collection, which significantly mitigates experimental bias and ensures minimal sample degradation. The required sample volume is 5 μL for each recording, and samples are categorized accordingly, with only one type per participant and without replicates or monitoring of treatment responses. The datasets herein are employed to provide a direct comparison with

extant literature and to verify generalizability for practical diagnostic applications. An additional cancer dataset includes measurements in tubed urine, supplemented by external cancer patient samples [8].

4.1. Data Sources

Two open-archive datasets were employed. First, brain-tumor blood was collected from 192 brain-tumor patients admitted to the National University Hospital, Singapore, under ethics approval number 2015/01203 [1]. Then, the Taiwan cohort was assembled by recruiting 167 patients with various cancers: 59 (35.3%) oral, 33 (19.8%) esophageal, 26 (15.6%) throat, 18 (10.8%) gastric, and 31 (18.6%) colorectal cancers [2]. For each patient, one blood sample was collected on the hospital admission day, prior to cancer treatment. Additional normal = 60 suspected-patients who had no established diagnosis and were confirmed short-term (Days to Weeks) negative were also recruited for comparison purpose.

4.2. Data Preprocessing Steps

Data were manually cleaned to remove abnormal points, operations and noise. Then the spectra were denoised through wavelet analysis. An algorithm was applied to identify the baseline in the spectrum, which was then removed. The spectra were calibrated and normalized before data analysis. These processes are crucial as spectra collected from Raman instruments are usually complex, extensive, and noisy. Creating a robust analysis method requires attention to variability in the samples, whether due to different donors or sample preparation inconsistencies. Electrical noise can introduce artificial peaks, complicating model building and potentially obscuring key features [9]. The quality of Raman based diagnosis also depends on several preprocessing stages that improve data quality while preserving chemical information. Such strategies explore the exclusion of outliers followed by application of various corrections including baseline, smoothing, normalization, and scaling [10].

4.3. Feature Extraction Methods

Feature extraction is a fundamental task in machine-learning analysis for multiplexed disease predictive-model construction, involving two strategic approaches. The more straightforward raw-signal approach involves directly utilizing molecular fingerprinting data as input for predictive models; the database-tuned approach implements pre-extracted features by integrating prior knowledge gleaned from existing databases, thereby capturing characteristic disorder-related information absent in raw signals. The raw-signal strategy employs time series feature extraction methods such as Symbolic Aggregate Approximation (SAX) and Template-Based Matching (TBM), each catering to distinct pattern types and enabling a comprehensive representation when combined. The feature extraction workflow commences by splitting each time series signal into differently sized and overlapped subseries using a sliding window method, with each subseries subsequently converted into feature vectors by SAX and TBM for further analysis.

The SAX technique transforms a real-valued time series into a symbolic representation via three steps: normalization, dimensionality reduction through Piecewise Aggregate Approximation (PAA), and discretization based on predefined breakpoints corresponding to a desired alphabet size. Optimal window lengths for SAX application are determined by matching the scattering profile of SAX features to the intrinsic properties of the fingerprinting signals, with these parameters derived through physics-driven spectral analysis and meticulous mapping procedures. TBM is employed to extract template features that model the temporal shapes of specular-reflection signals within the mercury phosgene molecular fingerprinting (MPF) waveform. This method excels at representing signals with well-defined onset and offset characteristics and monotonic amplitude profiles; it operates by constructing templates for each temporal feature and computing the distance between each subseries and these templates, facilitating the derivation of similarity-based feature vectors [7] [1].

5. Model Development

Trained machine learning models peripherally facilitate the labeling of molecular fingerprints, thereby bridging raw feature data and practical disease-monitoring applications. In the case of two-dimensional nuclear magnetic resonance (2D NMR) spectroscopy, cholesterol estimation exemplifies the transformation from Ni chemical shifts to clinically relevant information [1]. The training pipeline combines data preparation, normalization, and augmentation to induce generalization, utilizing three models for cholesterol and lipoprotein profiling: artificial neural network (ANN), support vector machine (SVM), and random forest (RF).

Molecular fingerprinting methods inevitably confront practical questions of data availability for model training, heterogeneous data sources across screening programs, and time or budget constraints that may preclude extensive collection or laboratory assays. Hence, a data assimilation step synthesizes and harmonizes available observations for coherent machine learning [8]. Moreover, the General Architecture for Text Engineering (GATE) framework ingests, standardizes, and manages clinical and textual data interchangeably, developing rule-based systems to optimize workload through classification and annotation of incoming documents [11]. From a dataset perspective, two emerging challenges entail realistic collection modeling and direct identification of incompleteness in samples.

Subject to rigorous preprocessing and feature extraction, acquired data then transits to supervised model development capable of accurate, generalizable predictions. The pipeline initiates with the derivation of fingerprint vectors, followed by the training of a remote analysis model, subsequently undergoing validation with field- and laboratory-sourced samples. An emerging task involves concurrent training across multiple collections differing in source or distribution, while another models the temporal progression of incomplete samples.

5.1. Training the Machine Learning Models

The machine learning models are trained on measured molecular profiles and clinical test results or medical records. Measurements of 2D molecular profiles each take only a few minutes. A single profile contains an ensemble of molecular information that enables rapid and automated blood analysis. Multiplex screening models can be deployed on a separate clinical platform to identify specific diseases. The pipeline thus provides efficient tools to accelerate screening and decision-making workflows. Each model allows profile dimensionality reduction and rapid classification under a given clinical subject or disease [1].

5.2. Validation Techniques

Overfitting constitutes a significant obstacle in high-dimensional data analysis. Cross-validation, a resampling method, addresses this issue by reusing the dataset partitioned into training and testing subsets. The model is trained on the training data and assessed using the validation data through performance metrics such as accuracy, sensitivity, and specificity. Random forest—an ensemble Machine Learning technique known for efficiency with tabular data—utilises out-of-bag (OOB) samples, inherently providing an unbiased estimate of the generalisation error [3]. Conversely, Support-Vector Machines (SVMs) require an additional layer of validation to tune hyperparameters, often achieved through k-fold cross-validation by splitting the training set into k smaller sets [11].

5.3. Performance Metrics

A series of indicators were employed to provide a multidimensional view of the ML model's predictive skills. Instead of computing multiple performance indicators for each model and algorithmic configuration, three were selected: accuracy, sensitivity, and specificity; and the sole results of the best-performing AutoML framework. With a COVID-19 incidence in Malaysia at 4.62 % in January 2022, and an average sample submission rate of 225 samples per week, an epidemiological baseline recall level was established at 0.954 [1]. The best-performing AutoML

framework in the study exceeded this epidemiological baseline recall level significantly. The data is more suitable for sensitive detection of a mild stage of LD rather than a multi-class classification to gauge the severity of LD. The dataset is reasonably balanced, but the performance of each model depends on a trade-off between specificity and sensitivity.

The performance indices are defined as follows:

Accuracy: The ratio of correctly predicted samples to all the data samples

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Sensitivity: Measures the ability of the model to correctly detect the positive (disease) cases

Sensitivity = TP / (TP + FN)

Specificity (true-negative rate): Measures the ability of the model to accurately reject the negative (healthy) cases

Specificity = TN / (TN + FP)

where TP and TN, respectively, signify the true positive and true negative samples, and FP and FN represent the false positive and false negative ones. Sensitivity allows capturing the most positive samples and provides a high confidence level in negative predictions, which is desirable in medical diagnosis.

6. Results

To demonstrate the feasibility of disease screening by molecular fingerprinting of blood samples, different models were constructed and evaluated to perform multiplexing tasks of screening diverse disease states based on the input molecular fingerprints of blood samples. A range of conditions were considered, summarised in Table 1. The number of blood samples used to represent each condition is also documented in the table. For screening of specific conditions, a binary classification model was built to differentiate between healthy and affected individuals. For simultaneous screening of multiple disease states, a multi-class strategy was employed, and the model was tasked with the identification of the corresponding condition based on the molecular fingerprint. During model training, a 20% subset of the data was randomly drawn for independent validation. Model performance was quantified using the area-under-curve (AUC) metric.

6.1. Model Performance Overview

The model's performance is assessed by considering five disease categories alongside asymptomatic controls. The category of symptomatic controls includes all negatives that originate from symptomatic individuals exhibiting symptoms indicative of an acute infection.

6.2. Comparison with Traditional Methods

Clinical molecular fingerprinting of blood samples for rapid multiplexing and disease screening represents an efficient, high-throughput method for detecting pathogenic changes and biomarkers indicative of diverse conditions. By leveraging the information encoded in molecular fingerprints and applying trained machine-learning models, this approach facilitates the screening of multiple diseases concurrently from widely available blood samples. Consequently, the approach is particularly well suited for initial diagnostics, significantly reducing overall screening times, laboratory workloads, and associated costs. This performance advantage is established with publicly available, standardized clinical spectral sets.

Effective disease screening methods critically underpin global health networks. Traditional infectious disease diagnostics are predominantly culture-based, rendering them time- and labor-intensive [12]. Real-time polymerase chain reaction techniques, while more rapid, remain confined to single-pathogen tests. Recent developments in high-throughput, multiplex technologies and their integration with machine-learning models enable rapid diagnosis of

multiple diseases simultaneously. explored the combination of epidemiologic investigations and molecular genotyping, revealing the limitations of visual comparison and the need for automated computational systems to analyze large datasets of DNA fingerprints. Despite the availability of spectral blood-sample measurements, the application of molecular fingerprinting in conjunction with machine-learning models remains underexploited for multiplex clinical disease screening.

6.3. Case Studies

The multiplex disease-screening model was further investigated through a collection of test cases, spanning both the present and related studies [1]. Table 6.3 presents a catalogue of these cases, briefly summarizing the nature of the data and the objective of the machine-learning analysis.

Across all scenarios, both training and testing datasets were employed. Whenever possible, distinct or cross-validation datasets were deployed during testing; otherwise, unseen training data points served to evaluate predictive performance. Case 4, evaluated on a group of seven candidates, constituted a prospective testing set. Despite the small sample size, the limited population of healthy individuals with hyperlipidaemia (HLP) endorses the significance of the inference. Case 1 pertains to Schistosoma detection in human serum, with particular reference to the first multiplex diagnostic study conducted with the porous complex.

Relevant information: - Table 6.3 enumerates the various study cases along with the type of measured data and the target of investigation. - An additional numbered case corresponds to an analytical study of multiplexed screening concerning hyperlipidaemia (HLP), with detailed data not included in the present account [3]. - Simulation data serve for method development in multiplex Schistosoma detection within a porous complex (case 1).

7. Discussion

For acute diseases such as stroke, timely diagnosis and management are critical for patient outcomes [1]. Matrix-assisted laser desorption/ionization mass spectrometry profiling of human blood serum, integrated with machine learning, offers a rapid means to screen and identify various stroke subtypes. The classification system yields results within minutes, classifies beyond binary categories with notable sensitivity and specificity, and functions in a label-free modality. Accordingly, it demonstrates potential to expedite clinical decision making, guide interventions, and support in the early detection of stroke among patients presenting to emergency departments.

Mass spectrometry emerges as a fast, cost-effective, and increasingly accessible label-free technology for molecular screening. Although the direct inference of blood components based on molecular patterns presents challenges, machine learning facilitates inference and classification within systems comprising hundreds to thousands of analytes. This synergy paves the way for high-throughput and multiplex screening. Mass spectrometry-based molecular fingerprinting, when combined with AI, is notably suited for multiplex approaches and can be extended to encompass a broad spectrum of diseases. The present work establishes a foundation for further development in this domain [2].

7.1. Implications of Findings

The combination of molecular fingerprinting techniques and machine learning models presents a significant advance for rapid, multiplex disease screening from routinely collected blood samples. The analytical approach developed circumvents the need for specialized markers or primers, enabling sensitive and high-throughput classification of multiple blood-borne diseases from complex molecular profiles. Such a tool holds particular promise as a field-deployable screening platform, capable of facilitating differential molecular diagnoses and guiding subsequent molecular and confirmatory testing, especially in resource-limited settings. This capability aligns with prior observations on the critical importance of identifying multiple

pathogens simultaneously, as exemplified in polymicrobial bacteremia where early multiplex detection markedly influences patient management [13]. Beyond infectious diseases, molecular fingerprinting offers an extensive reservoir of biochemical and structural information that can be efficiently locked and interpreted by modern artificial intelligence, finding utility across a broad spectrum of medical contexts.

Although the present findings underscore the utility of this integrated approach when standardized pathogen panels are available, emerging data also suggest that advanced anomaly-detection frameworks have the potential to bridge gaps for unexpected or unknown agents. This capability provides a route to extend the technology's reach beyond the direct availability of training data. Accurately capturing the unique and stable molecular constituents of both infectious and non-infectious pathologies may expedite cascaded screening paradigms that more effectively incorporate other molecular and clinical data streams. Such multilevel classification systems would enable a more complete disease-characterization framework, not only identifying whether a patient currently suffers from an infection but also informing more precisely which additional screening tools or treatments should be prioritized [2].

7.2. Limitations of the Study

The present pilot study illustrates the potential of molecular fingerprinting, in conjunction with machine-learning algorithms, to facilitate multiplex screening of diseases in individuals unobtrusively and rapidly, with a low-cost, quick-turnaround tool requiring only a few drops of blood. This approach relies on widely accessible clinical instrumentation and readily available low-volume biological samples, thus enriching the range of existing molecular diagnostic methodologies.

Potential limitations include the fairly large volume of data generated by each spectrum. Although the present work employs wavelength-by-wavelength absorbance values, dimension-reduction strategies such as feature selection and principal-component analysis have been previously applied in the contexts of infrared spectrochemical and mass-spectrometric fingerprinting, respectively, for the characterization of biological samples, including early-stage Lyme disease, sepsis, and prostate cancer [14]. As a result of the cursory nature of the Raman-scatter spectra reported presently, a substantial number of data features are needed in the fingerprint to provide effective classification; nevertheless, the overall data volume is significantly smaller than its analogous spectrochemical forms.

7.3. Future Research Directions

Future investigations should consider expanding the scope beyond cancer detection toward other physiological and pathological states such as neurological disorders, cardiometabolic diseases, and inflammation. The adaptation of molecular phenotyping for broader application is warmly encouraged. As datasets become larger, the merit of deep learning methods becomes expected to surpass that of conventional techniques. More data will, therefore, serve to enhance integrated treatment, including feature selection, regression, clustering, and classification processes. However, the challenges of overfitting, model interpretability, and the black-box nature of deep models remain critical and must be addressed emphatically. Techniques such as regularization, dropout, and early stopping are indispensable to counter overfitting; explainable artificial intelligence approaches are crucial to elucidate hidden model properties; and the general difficulty of balancing model complexity and generalization stays inherently important. Given the availability of large datasets, further exploration of advanced algorithms including deep learning, gradient boosting, and transfer learning could be highly beneficial for identifying the latent biomarkers and complex relationships that are intimately correlated to physiological changes. The translation of integrated molecular fingerprinting and machine learning technology into clinical workflow will benefit from additional studies aimed at alleviating current challenges and broadening the array of malady types that can be effectively screened. [15] [2] [1]

8. Ethical Considerations

Clinical translation requires informed consent and maintains data privacy, in accordance with relevant legislation [16]. Researchers must adhere to national and international regulations regarding sample collection and storage. The system must record pathologies such as COVID-19 status, pneumonia, or thrombosis, and any additional healthcare data that may have influenced Raman results; this information remains within healthcare facilities. Model confidentiality mandates responsible publication, limiting the dissemination of detailed architectures and weights to authorized personnel.

8.1. Informed Consent

The study protocol for the molecular fingerprinting of blood samples utilized with machine learning for multiplex disease screening, as exhibited in [1] , was assessed by the University Institutional Review Board. Written informed consent conforming to the Declaration of Helsinki was procured from all participants [3].

8.2. Data Privacy

Multiple factors, including the sensitivity of data such as genomes and medical records, privacy concerns, and operational factors such as network interruption and long processing time, severely limit the utility, accessibility, and interoperability of genomic data [17].

Molecular fingerprints acquired with a two-dimensional (2D) NMR experiment and information extracted via machine learning models contain encrypted information of personal genomics and health data about individual patients within a single molecular signature. Proper data curation and information retrieval could facilitate the seamless transition of molecular profiles into real-world applications.

8.3. Ethical Use of AI in Healthcare

Regarding the ethical use of AI for multiplex screening of diseases involving human blood in AI systems, the study leverages a hybrid approach that integrates machine learning, molecular fingerprinting of blood samples, and analytical chemistry to maximize the strengths of each discipline while minimizing the weaknesses of any one approach. Plasma from healthy individuals and patients undergoing treatment for Type 2 Diabetes Mellitus generated molecular fingerprints by both Raman and Fourier-transform infrared (FTIR) spectroscopies, providing two very different, but complementary and confirmatory, data streams for predictive screening. Given the nonlinear character of the data, a deep-learning model was developed and implemented to combine the two molecular fingerprinting datasets and identify diabetic versus healthy individuals with an accuracy, precision, and recall >95%. The approach opens a path to the rapid, non-invasive, inexpensive, and multiplexed screening and monitoring of a wide variety of diseases and conditions using molecular fingerprinting coupled with machine-learning algorithms, which have the potential for immediate and global impact on the quality of healthcare. [18]

9. Applications

Blood-based prospective assays are attractive tools for rapid, noninvasive, and multiplex screening of major human illnesses. A generic molecular fingerprinting method based on two-dimensional (2D) T_1-T_2 correlational nuclear magnetic resonance (NMR) relaxometry is proposed—coupled with machine learning (ML) as a pre-trained molecular profile decoding system. Blood samples are optimally prepared by removing the interference from red blood cells for direct qualitative and quantitative fingerprinting of a variety of diagnostic markers including C-reactive protein (CRP) and COVID-19, liver and kidney functions, as well as early-stage cancers. At a proof-of-principle level, the results demonstrate that these markers can be detected at high sensitivity and specificity across a broad range of physiological conditions, as well as screened simultaneously without biochemical purification.

9.1. Clinical Diagnostics

Rapid discovery and assessment of emerging infectious threats is a recurrent problem for large populations. Conventional polymerase chain reaction (PCR)-based methods rely on a closed set of primers, limiting their capacity to anticipate the capacity of rapid and efficient surveillance. Molecular fingerprinting is a homogenized sample containing a heterogeneous molecular mixture with physicochemical information to render a wide variety of medical/biological information. Translating such fingerprints for relevant medical/biological information remains an unresolved challenge. The unique resonant spectrums derived from either nuclear magnetic resonance (NMR), mass spectrometry (MS), or infrared (IR) spectroscopy enable such fingerprints to be acquired rapidly [1]. Handheld NMR, MS, or IR devices further permit rapid fingerprinting in the community of interest. Consequently, recording the molecular fingerprint coupled with advanced machine-learning algorithms provides an ideal solution to perform large epidemiological surveys and timely interventions.

Studies conducted under rigorous clinical control using either multi-chemical 2-dimensional (2D) NMR, MS, or IR fingerprinting coupled with artificial intelligence (AI) successfully demonstrated core capabilities in rapidly differentiating between healthy and diseased states, screening solid and liquid biopsies for pathologies, predicting various pathological phenotypes, and performing temporal disease progression. Given the standard of care involved with performing blood draws in population health screening studies, molecular fingerprinting typically focuses primarily on blood assessments. [19][20]

9.2. Public Health Monitoring

Molecular fingerprinting techniques for blood samples and machine learning classification models enable multiplexed disease screening with single samples. By coupling molecular fingerprinting with machine-learning methods, key information related to multiple diseases can be extracted from blood samples and used to build classification models for disease detection.

Multiple combination-choice scenarios involving organism, antibiotic, and disease, as well as a cartridge capable of multiple analyte extraction, further enhance this approach. Applications encompass the detection of bacterial/viral infections, antibiotic resistance identification, and diagnosis of diseases such as malaria, tuberculosis, and sepsis.

9.3. Personalized Medicine

The combination of molecular fingerprinting of blood samples and machine learning enables multiplex disease screening in a personalized fashion [21]. Digital phenotyping of blood, facilitated by molecular fingerprints, offers an alternative to the analysis of gene-variant and gene-expression data as a means of carrying out such personalized medicine [22]. Nonetheless, rapid, high-throughput, and label-free molecular phenotyping of blood remains a challenge [1].

10. Conclusion

Molecular fingerprinting of blood samples—an approach that integrates fingerprinting techniques with machine learning—enables rapid disease screening on highly portable platforms. By capturing the interplay of multiple molecular processes among diverse biomolecules in the form of a single spectral pattern, an effective disease-specific spectroscopic fingerprint can be generated. With a state-of-the-art feature-extraction and machine-learning algorithm, such a molecular fingerprint can be decoded efficiently to screen a wide range of diseases from a single blood sample. The technique demonstrates an exceptional ability to characterize molecular changes related to cancer and Diabetes Mellitus, discriminating effectively between healthy, benign, and malignant subjects. Predicated on robust machine-learning models, the method offers rapid, high-throughput, and label-free molecular phenotyping of blood, thereby holding great promise for in vitro disease diagnosis and monitoring. The technology facilitates frequent testing via minimally invasive liquid-biopsy read-outs, with an overall experimental time under 6

minutes ensuring both high sensitivity and spectral resolution.

References:

1. W. Kung Peng, T. T. Ng, and T. Ping Loh, "Machine learning assistive rapid, label-free molecular phenotyping of blood with two-dimensional NMR correlational spectroscopy," 2020. ncbi.nlm.nih.gov
2. L. Liu, X. Chen, O. Olayemi Petinrin, W. Zhang et al., "Machine Learning Protocols in Early Cancer Detection Based on Liquid Biopsy: A Survey," 2021. ncbi.nlm.nih.gov
3. L. Mathot, E. Falk-Sörqvist, L. Moens, M. Allen et al., "Automated Genotyping of Biobank Samples by Multiplex Amplification of Insertion/Deletion Polymorphisms," 2012. ncbi.nlm.nih.gov
4. Y. Hu, "Molecular Techniques for Blood and Blood Product Screening," 2018. ncbi.nlm.nih.gov
5. H. Skutkova, M. Vitek, M. Bezdicek, E. Brhelova et al., "Advanced DNA fingerprint genotyping based on a model developed from real chip electrophoresis data," 2019. ncbi.nlm.nih.gov
6. T. M. Ghazal, H. Al Hamadi, M. Umar Nasir, undefined Atta-ur-Rahman et al., "Supervised Machine Learning Empowered Multifactorial Genetic Inheritance Disorder Prediction," 2022. ncbi.nlm.nih.gov
7. O. Tutsoy and G. Gul Koç, "Deep self-supervised machine learning algorithms with a novel feature elimination and selection approaches for blood test-based multi-dimensional health risks classification," 2024. ncbi.nlm.nih.gov
8. Z. Yaari, Y. Yang, E. Apfelbaum, C. Cupo et al., "A perception-based nanosensor platform to detect cancer biomarkers," 2021. ncbi.nlm.nih.gov
9. D. L Tong, D. J Boocock, C. Coveney, J. Saif et al., "A simpler method of preprocessing MALDI-TOF MS data for differential biomarker analysis:stem cell and melanoma cancer studies," 2011. [PDF]
10. D. L Tong, D. J Boocock, C. Coveney, J. Saif et al., "A simpler method of preprocessing MALDI-TOF MS data for differential biomarker analysis: stem cell and melanoma cancer studies," 2011. ncbi.nlm.nih.gov
11. L. Zhang, Q. Liu, Y. Guo, L. Tian et al., "DNA-based molecular classifiers for the profiling of gene expression signatures," 2024. ncbi.nlm.nih.gov
12. H. Salamon, M. R Segal, A. Ponce de Leon, and P. M Small, "Accommodating error analysis in comparison and clustering of molecular fingerprints.," 1998. ncbi.nlm.nih.gov
13. D. Ortiz Velez, H. Mack, J. Jupe, S. Hawker et al., "Massively parallel digital high resolution melt for rapid and absolutely quantitative sequence profiling," 2017. ncbi.nlm.nih.gov
14. B. T Luke and J. R Collins, "Examining the significance of fingerprint-based classifiers," 2008. ncbi.nlm.nih.gov
15. T. Beduk, D. Beduk, M. Rahil Hasan, E. Guler Celik et al., "Smartphone-Based Multiplexed Biosensing Tools for Health Monitoring," 2022. ncbi.nlm.nih.gov
16. N. Scudder, D. McNevin, S. F. Kelty, S. J. Walsh et al., "Forensic DNA phenotyping: Developing a model privacy impact assessment," 2018. [PDF]

17. M. Robinson and G. Glusman, "Genotype Fingerprints Enable Fast and Private Comparison of Genetic Testing Results for Research and Direct-to-Consumer Applications.," 2018. [PDF]
18. P. Kumar Dabla, "Unlocking new potential of clinical diagnosis with artificial intelligence: Finding new patterns of clinical and lab data," 2024. ncbi.nlm.nih.gov
19. H. Bangalore, "Leveraging DNA Fingerprinting to Combat Sample Contamination in High-throughput Molecular Testing," 2023. researchgate.net
20. M. Huber, K. V. Kepesidis, L. Voronina, M. Božić, et al., "Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring," *Nature*, vol. 2021. nature.com
21. W. DeGroat, H. Abdelhalim, K. Patel, D. Mendhe et al., "Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine," 2024. ncbi.nlm.nih.gov
22. Z. Ahmed, "Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis," 2020. ncbi.nlm.nih.gov