**BCDA**

**Generating Predictive Models from Other Large Datasets**

* TCGA - The Cancer Genome Atlas

TCGA is by far the most comprehensive publicly available compilation of tumor profiles and includes a large number of data types spanning :

* Genomics refers to the study of an organism's complete set of DNA, including all of its genes.
* Epigenomics involves studying the changes in gene expression that are not caused by alterations in the DNA sequence itself.
* Proteomics is the study of all the proteins produced by an organism or a specific cell type.
* Histopathology is the microscopic examination of tissue samples to study the changes in cell structure and organization that occur in diseases like cancer.
* Radiology

Mature technologies have led to the production of a wide-ranging array of datasets. Some of these technologies are:

1. DNA methylation: DNA methylation is a process where methyl groups are added to the DNA molecule, to study the change in patterns. Machine learning algorithms can be trained on DNA methylation profiles to improve prediction accuracies for cancer classification.
2. **Large-scale proteomics studies** involve the analysis of a vast number of proteins simultaneously. In the context of cancer research and precision medicine, large-scale proteomics studies aim to understand the complex protein networks and alterations that occur in cancer cells.
3. **Perturbation studies** include the study of change based on cell viability or cytotoxicity assays using small molecules, RNAi or CRISPR screens, and protein-protein interaction networks.

* Integrating multiple data types improves prediction accuracy.

**Data Quality and Model Selection Are Key**

Strategies like cross-validation, increasing training set size, and using ensemble approaches help mitigate risks of overfitting.

Overfitting occurs when a model learns from noise or artifacts instead of the true signal, leading to poor generalization of unseen data. It is essential to inspect and correct data for inaccuracies, missing values, biases, and use techniques like manual curation of features to address overfitting risks.

Performance of ML model is commonly measured using AUROC(Area under receiver operating characteristics Curve) or AUC to balance sensitivity and specificity for optimal classification in different applications.

AUC >0.80 is considered good, but its clinical acceptability varies based on use. AUC evaluates model performance at a population level, not individual predictions. AUPRC (Area under precision recall curve) is preferred for imbalanced datasets focusing on accurately detecting the positive class.

Routine maintenance of ML model is essential to ensure performance remains stable over time by addressing concept drift( where the relationship between input and output variables changes unexpectedly).

**EARLY DETECTION, DIAGNOSIS, AND STAGING OF CANCER**

AI adds the advantage of scalability and automation to these key aspects of cancer treatment.

* DNNs are applied to large images like "H&E Stained WSIs" (Hematoxylin and eosin stain for whole slide imaging) of tissues derived from biopsies and surgical resections, for classifying cancer cells from healthy ones. (AUC > 0.99). They are also used for more complicated cancer subtype classifictaion. e.g such as

adenocarcinoma vs. adenoma in gastric and colon cancers and

adenocarcinoma vs. squamous cell carcinoma in lung tumors

* An inception-v3 architecture based model, "DeepPATH" was developed and applied for concurrent classification of 3 classes: normal, lung adenocarcinoma and lung squamous cell carcinoma. (AUC = 0.97).

DNNs can be applied to images acquired through non-invasive techniques such as CT scans, MRI, mammograms and photos of suspicious lesions.

Examples of Deep Learning in Cancer Diagnosis:

1. Skin Cancer Classification: Deep learning model (Inception-V3) trained on images of various skin lesions. Achieved higher accuracy (AUC: 0.91-0.94) in classifying melanoma and carcinoma compared to dermatologists. Robust to variations in image quality (camera angles, exposure).

2. Lung Cancer Detection: Deep learning model built using CT scans of patients with lung disease. Effectively classified textural patterns like ground-glass opacity and micronodules (average accuracy: 0.85).

3. Gastric Cancer Metastasis Prediction: Deep learning model developed using CT scans to predict occult peritoneal metastasis. Achieved higher AUC (0.92-0.94) compared to traditional methods (AUC: 0.51-0.63).

4. Prostate Cancer Classification: Deep learning model trained on MRI images to differentiate prostate cancer from benign conditions. Reported AUC of 0.84 in distinguishing cancerous and non-cancerous cases.

5. Breast Cancer Risk Assessment: Ensemble approach with three independent deep learning models used to predict cancer risk from mammograms. Achieved AUC of 0.75-0.88 and improved specificity (1.2%-5.7%) and sensitivity (2.7%-9.4%) compared to radiologists.

AI-based smartphone apps are being developed to detect cancerous lesions conveniently outside hospitals. However, there is a concern that these apps might miss cancers (false negatives) leading to delayed diagnosis and treatment.

**Cancer Staging and Grading**

* In prostate cancer, the "Gleason score" is used for staging. The Gleason score is a combination of two scores measuring the prevalence of tumor cells in two distinct locations on a slide.
* WSI and H&E stained prostatectomy specimens were used to train and test a DNN(inception v-3) and k-nearest neighbour classifier based model to predict Gleason scores.
* The model turned out to be more accurate(0.70) than a panel of 29 pathologists(0.61).
* Staging using radiology was done using a DL approach based on SENet and DenseNet to predict liver cancer grade from MRI images. (AUC = 0.83)

ML is required for the classification of multidimensional data obtained fro next get sequencing methods.

Omics technologies and machine learning algorithms have advanced, with studies showing the effectiveness of random forest classifiers in predicting hard-to-diagnose central nervous system cancer subclasses based on DNA methylation profiles. (The subclass predictions did not match pathologists’ diagnosis, but follow-up of those select cases revealed that approximately 93% of those mismatched cases were in fact accurately predicted by the model.)

Deep learning methods like DNN achieve high accuracy in distinguishing healthy and tumor tissues(AUC = 0.94) but face challenges in multiclass cancer type classification (AUC = 0.70) due to intratumor heterogeneity and shared mutations.

**On the Road to Early Cancer Detection Using AI**

Liquid biopsies like ctDNA(circulating tumor DNA) or cfDNA, enable monitoring of cancer risk, treatment guidance, and predicting MSI status(microsatellite instability status) in endometrial cancer.

Machine learning models like Lung-CLiP(cancer likelihood in plasma) predict the presence of ctDNA in blood sample.(AUC = 0.69-0.98)

Random forest classifiers trained on cfDNA fragment sizes are being developed to predict ctDNA presence in blood samples (AUC = 0.91–0.99).

CancerSEEK is a comprehensive blood test for eight cancer types, detecting early cancer and predicting cancer types using mutations and protein expression levels directly from ctDNA. This work is particularly important because five of the eight cancer types covered in this test have no early screening tests currently available.

**DETECTING CANCER MUTATIONS USING MACHINE LEARNING**

AI is being used to detect key mutations in cancer from histopathology images, offering a cost-effective and faster alternative to NGS.

Deep learning models like DeepPATH and transfer learning approaches can identify 6 key mutations: STK11, EGFR, FAT1, SETBP1, KRAS, and TP53 from various cancer types using image data.

AI models like ResNet18 and MSINet have shown promising results in detecting MSI status in tumors from histopathology slides, with AUCs ranging from 0.75 to 0.93.

Tumor mutation burden (TMB) estimation from histopathology slides is an active research area, with models like Image2TMB and GoogLeNet showing high accuracy in predicting TMB status.

Chromosomal instability prediction from histopathology slides is also being explored as a driver of cancer evolution.

Validation of AI models for mutation detection may vary based on sequencing platforms, tissue preservation techniques, and ethnicities of the cohorts used for training.

**Determining Tumor Cells of Origin**

Determining the cell of origin of tumors for personalized therapies, with site-specific treatments shows better efficacy than systemic chemotherapies. The tissue of origin is traditionally identified using IHC and gene expression profiling, but new AI-based models using mutation counts can predict origin with 83-91% accuracy.

AI methods are enhancing cancer mutation detection, aiding in understanding cellular mechanisms, screening patients for specific mutational profiles, and designing targeted therapies.AI is expected to evolve in predicting the functional impact of mutations, analyzing noncoding mutations, and detecting mutations from histopathology images for clinical relevance.

AI advancements may help predict therapy resistance, mutation status changes, and tumor evolution directly from histologic pattern changes in pathology images.

**CHARACTERIZING THE TUMOR MICROENVIRONMENT**

AI approaches in digital pathology, while highly accurate, are often considered "black box" as they lack interpretability in explaining disease classifications compared to pathologists who rely on well-documented features and training. AI can automate tasks like estimating tumor cellularity, crucial for assessing residual disease and selecting tissue blocks for further analysis, reducing subjectivity and laboriousness in traditional methods.

DNN was used to quantify tumor cellularity directly from H&E-stained WSI, showing good concordance with pathologists' reports, but larger datasets are needed for further validation.

AI is also used to assess the tumor microenvironment (TME) spatially and quantitatively, aiding in understanding tumor evolution, metastasis, and response to therapies. The identification and quantification of lymphocyte infiltration using DNN has resulted in achieving high accuracy in classifying TIL status and reducing false positives.

DNNs like VGG-Net and autoencoder-based models are utilized to distinguish benign from malignant tissues and classify stroma in histopathology images with high accuracy.Multiplexed imaging platforms and advanced network architectures like GoogLeNet are increasingly used to study complex cell interactions within the TME.

DNNs are also employed to deconvolve bulk RNA-seq profiles into cell types based on scRNA-seq data, with efforts to expand single-cell profile availability through initiatives like The Human Cell Atlas.

Understanding individual cell types in the TME, especially immune cells and neoantigens, is crucial for successful checkpoint immunotherapy and personalized treatments. Neoantigens, arising from tumor-specific mutations, play a key role in antitumor immune responses and are studied for their potential in immunotherapies like personalized peptide vaccines.

Tools like NetMHC, based on artificial neural networks, predict neoantigens by analyzing mutated peptides' binding affinities to MHC class I alleles, with newer tools incorporating advanced AI algorithms like NLP and CNNs.

A patient-specific approach using NLP models to predict neoantigen sequences directly from a patient's MS data was also developed, showing promise for personalized peptide vaccine development but requiring further validation.

**DISCOVERY OF THERAPEUTIC TARGETS AND DRUGS**

* AI models integrate diverse datasets for drug target prediction, such as clinical data with gene expression profiles, leading to the identification of potential drug targets in various cancers.
* AI is used in drug design to generate molecules with desired properties using techniques like reinforcement learning and graph convolutional networks.
* Generative adversarial networks (GAN) and reinforcement learning are applied for molecule generation tasks, achieving high performance in properties like drug likeness and solubility.
* Drug repurposing, a cost-effective alternative, is accelerated by AI using rich transcriptional datasets like LINCS to identify new therapeutic uses for existing drugs.
* AI models trained on drug-perturbed transcriptional profiles predict therapeutic use categories for drugs and prioritize repurposing candidates based on chemical structural similarity.
* Publicly available datasets from cell viability assays are utilized to train AI models for drug repurposing efforts, predicting effective drugs for individual patients based on somatic mutation profiles.
* Various AI models like CDRScan and DeepDR predict novel drug-disease connections and repurposing candidates, leading to promising repurposing predictions in cancer and other diseases.

**PATIENT PROGNOSIS AND RESPONSE TO THERAPY**

Prospective patient identification for tailored therapies can reduce poor outcomes and high treatment costs, especially in immunotherapies with low response rates.

AI applications in oncology include logistic regression and XGBoost models to predict responses to immunotherapies like PD-1 inhibitors and checkpoint inhibitors.

Deep learning models like DNNs and CNNs are used to predict patient prognosis and response to cancer treatments based on omics data and imaging. Machine learning models can predict patient sensitivity to chemotherapy, targeted therapy, and immunotherapy using features from CT scans.

AI can assist in monitoring cancer progression and therapy response by analyzing pathology or radiology images, especially challenging in immunotherapies with atypical disease patterns.

Machine learning models, like CNNs, can predict response to neo-adjuvant chemotherapy in breast cancer patients using PET/MRI scans. Longitudinal CT scans of lung tumors can be analyzed using CNN with recurrent neural networks to predict overall survival after chemoradiation.

CURATE.AI offers personalized drug dosage adjustments for individual patients using dynamic patient-specific data points collected over time

**PREDICTING DRUG EFFICACY AND SYNERGY**

Machine learning algorithms are used to predict drug efficacy based on molecular features, utilizing large cancer drug efficacy datasets from cell line experiments.

Preprocessing is essential to minimize noise in datasets, such as cell line authentication and validation of in vivo data. Random forests and deep learning, DNNs, are commonly used for drug response prediction, with deep learning showing promise in identifying drug-cancer pairings and patient cohorts based on sensitivity.

Deep learning methods lack interpretability, but efforts like DrugCell aim to address this by developing interpretable models that resemble known biological processes.

Combining genomic data with other features like chemical information and biological data can enhance the prediction of single-drug or combination efficacy. Models integrating genomic features with chemical structures or biological interactions tend to produce higher accuracies in predicting drug synergy.

**CURRENT CHALLENGES AND FUTURE PERSPECTIVES**

* Data biases exist in AI models due to the underrepresentation of diverse populations in training datasets, leading to challenges in developing robust and inclusive models.
* Sharing well-documented code for AI models is crucial for transparency, reproducibility, and clinical relevance, but universal adoption is lacking.
* Electronic health records (EHR) are an underutilized data source in AI cancer models, requiring standardization and user-friendly tools for effective utilization.
* Building trust in AI-assisted decision-making among clinicians is essential, with a focus on quantifying uncertainty in models to improve confidence.
* Increasing model interpretability through research on deep learning mechanisms can enhance clinicians' acceptance of AI models in clinical settings.
* AI's future application in cancer care may shift towards prevention, utilizing data from genetic tests, EHR, wearable devices, and lifestyle factors for personalized risk assessment and early intervention.
* AI systems could remotely monitor cancer patients, provide personalized prevention and treatment plans, and assess cancer risk in real time based on genetic, environmental, and lifestyle factors.