Uterine Corpus Endometrial Carcinoma Prediction from Genomic Analysis with Machine Learning

1st Marina Adriana Mercioni
Politehnica University Timisoara
Timisoara, Romania
marina.mercioni@upt.ro

2nd Stefan Holban
Politehnica University Timisoara
Timisoara, Romania
stefan.holban@cs.upt.ro

***Abstract*— Endometrial carcinoma (EC) is a common uterine malignancy that still contributes significantly to cancer-related morbidity and mortality. EC identified in an advanced stage has a poor treatment response. In the United States, the most frequent cancer of the female reproductive organs is endometrial carcinoma. The American Cancer Society's estimations for uterine cancer in the United States in 2024 are approximately 67,880 new cases of uterine cancer will be diagnosed and approximately 13,250 women will die from uterine cancer. EC is the second most common gynecological malignancy in the world and the top in continental Europe. The clinically used EC diagnostic techniques are expensive, time-consuming, and not available to all patients. The fast expansion of computational biology has sparked significant research interest from both data scientists and oncologists, resulting in the creation of quick and cost-effective computer-aided malignancy monitoring systems. Machine learning (ML), a subset of artificial intelligence, enables drug development, early disease detection, successful therapy, and treatment modality selection. The use of ML methods in EC diagnosis, treatment, and prediction may be very important. This review provides a summary of EC, as well as risk factors and diagnostic procedures, before delving into a complete genetic investigation of prospective ML modalities for EC prevention, screening, detection, and prognosis.**

Keywords*— *algorithm, endometrial carcinoma, machine learning, malignancy mitigation, oncology, prediction, trends

I. Introduction

In 2024, the US is expected to have 2,001,140 new cancer cases and 611,720 cancer deaths. Cancer mortality has continued to fall until 2021, preventing almost 4 million deaths since 1991 due to smoking cessation, faster identification of certain malignancies, and improved treatment choices in both the adjuvant and metastatic settings [1]. Endometrial cancer is the fourth most common kind of gynecological cancer in Romania, accounting for 7.8%. Endometrial cancer ranks fourth among the causes of genital cancer death (5.7%-5.9%) [2]. Women with early detection or EC with a lower risk have a better prognosis. Individuals with higher stage EC who have developed recurrence have a worse 5-year survival rate, ranging from 47% to 58% for stage III EC patients and 15% to 17% for stage IV EC patients, and have fewer available prognostic or therapeutic options[3]. Expensive screening and a high rate of misdiagnosis contribute significantly to high illness fatality [4-5]. Endometrial cancer diagnosis and management are difficult and complicated, necessitating the competence of multidisciplinary team members who are experienced with all aspects of its examination and treatment. High body mass indices (BMI), as well as diabetes type II and insulin resistance, anovulation, menstrual disruption, amenorrhea, and infertility, have been linked to an elevated risk of low-grade EC and, in newer research, high-grade EC [6]. EC is twice as prevalent in overweight women and more than triple the risk in obese women. In a large epidemiologic study, the risk of endometrial cancers, both type 1 and type 2 carcinomas, was found to decrease with increasing age of first childbirth, by 11% overall, and the risk for women who first gave birth after 40 years was 44% lower than women who gave birth before the age of 25 [7-12]. Breastfeeding decreased the risk of endometrial cancer by 11% [13-14]. Based on existing meta-analyses, the World Cancer Research Fund (WCRF) found that coffee drinking likely protects against EC [15]. Analyses show that night shift job and sleep length are not significant risk factors for endometrial cancer in postmenopausal women [16]. Recently, the circadian rhythm has been demonstrated to be connected with EC, with the severity of EC being linked to night work and rhythm problems. As a result, circadian rhythm disorders (CRDs) might be one of the metabolic illnesses causing EC. Clock genes (CGs) govern circadian rhythm changes, which are further regulated by non-coding RNAs (ncRNAs). More significantly, the mechanism of EC induced by ncRNA-mediated CRDs is rapidly becoming clearer [17]. Lack of rhythmic regulation has been anticipated to result in uncontrolled growth and malignancy. This hypothesis is supported by research showing that circadian disturbance caused by low light at night or persistent jet lag increases tumor development [18]. According to recent findings, alcohol drinking may raise the risk of EC, although coffee and tea consumption may minimize it [19-21]. Because EC is associated with overweight and obesity, keeping a healthy body shape through a nutritious diet and exercise is the most significant strategy most women can take to lower their chance of developing EC. Women with EC have a higher risk of all-cause death when their BMI is higher than 40 [22-23]. The differential diagnosis faced while evaluating the most frequent presenting symptoms and indications of EC, abnormal vaginal bleeding and/or pelvic masses, vary from benign localized lesions to systemic illnesses and malignancies. Patients with endometrial malignancies have a variable prognosis based on

979-8-3503-4929-0/24/$31.00 ©2024 IEEE

several factors such as the histotype, size, grade, and comorbidities, in addition to the surgical-pathological staging's assessment of the disease's extent [24]. Although the Cancer Genome Atlas (TCGA) project has made significant progress in understanding the biological heterogeneity of EC, it is still unclear how best to apply molecular categorization in relation to adjuvant treatment, surgical staging, and surveillance scheduling. Devoted efforts are needed to more precisely define host factors, such as microbiome composition and the impact of BMI, and a deeper understanding of the molecular and immunological drivers of response and resistance to emerging therapies is crucial for the best possible design of next-generation studies, in order to manage EC comprehensively. Significant copy number abnormalities, little DNA methylation changes, low levels of the oestrogen receptor and progesterone receptor, and frequent genetic mutations were seen in uterine serous tumors and around 25% of high-grade endometrioid tumors [25-26].

II. Materials and Methods

A. Dataset

The Tumor Cancer Genome Atlas-Uterine Corpus Endometrial Carcinoma dataset was downloaded from the Genomics Data Commons Data. More cancer types were analyses within genome project, analyzing individual genome data. The focus is on **U**terine **C**orpus **E**ndometrial **C**arcinoma (UCEC), its length is 529 patients. The categories included in the data are biospecimen, clinical, copy number of variation, DNA methylation, proteome profiling, sequence reads, simple nucleotide variation, structural variation, transcriptome profiling, somatic structural variation.

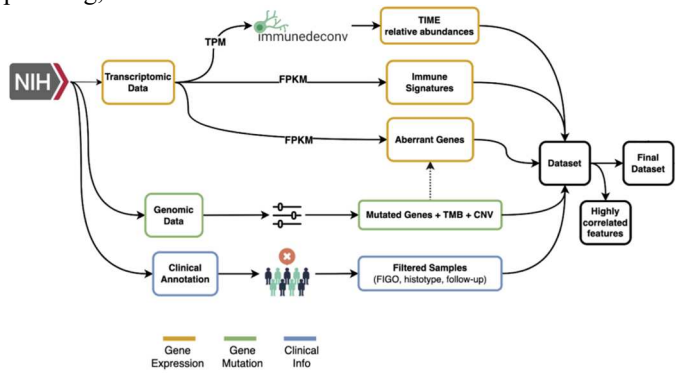

Fig. 1. Dataset overview [27].

According to the Fig. 1 the immune system impact has been analyzed in transcriptomic data, its effect, and aberrant genes. Adding to it, genomic data and clinical annotation. The features of the dataset such as "*Diagnosis Age*", "*Aneuploidy Score*", "*Subtype*", "*Patient Weight*", "*Disease-specific Survival status*", "*Fraction Genome Altered*", "*Neoadjuvant Therapy Type Administered Prior To Resection Text*", "*Mutation Count*", "*Overall Survival Status*", "*Radiation Therapy*", "*Tumor Type*" are only some features of the dataset.

B. Architecture

Two Machine Learning (ML) approaches (Histogram Gradient Boosting Classifier (HGB) and LightGBM Classifier (LGBM)) (Fig. 2), which are strong algorithms known for their high accuracy in classification tasks, were used to evaluate each sample's probability of living without disease. Due to their efficiency and speed optimisations, HGB and LGBM are good options for analysing huge datasets that are often used in medical research. Faster model training and prediction are made possible by this efficiency, which might be advantageous in clinical contexts where prompt diagnosis is crucial. Predicting UCEC frequently entails assessing high-dimensional data, including genetic data. Numerous characteristics may be included in the prediction model since HGB and LGBM can both handle high-dimensional data successfully and efficiently.

In the context of UCEC prediction, high accuracy is crucial for identifying potential cases of endometrial carcinoma accurately. Balanced accuracy and precision measures were used to assess the model performances based on confusion matrices.

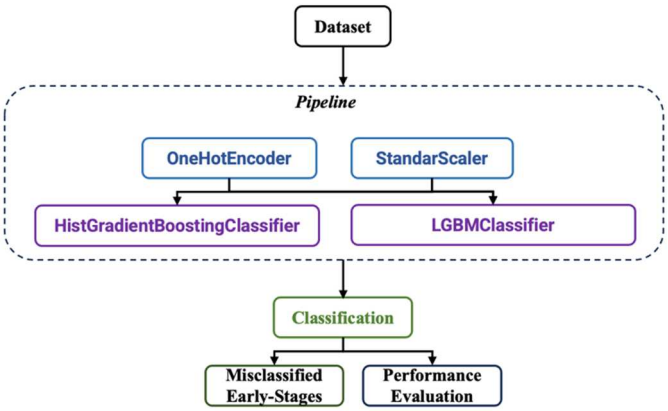

Fig. 2. Architecture overview.

Tree boosting is a popular and very successful ML technique. A scalable end-to-end tree boosting technique called *XGBoost* is frequently utilised to produce cutting-edge outcomes on a variety of ML tasks. It is a weighted quantile sketch for approximation tree learning and a unique sparsity-aware technique for sparse data. More significantly, it offers information on data compression, sharding, and cache access patterns to help construct scalable tree boosting systems. Combining these discoveries allows *XGBoost* to operate on billions of instances while consuming a fraction of the resources of other systems [28]. Native support for missing data (NaNs) is available for *XGBoost*. Based on the potential gain, the tree grower determines at each split point during training which samples with missing values belong to the left or right child. Samples with missing values are subsequently assigned to the left or right child when making predictions. Samples with missing values are mapped to the child with the most samples if no missing values were found for a particular feature during training. *LightGBM* achieves great accuracy while accelerating the training process by up to 20 times [29].

III. Results

A. Feature selection and Interpretation

There are multiple features in the dataset (62). The most prevalent features are as follows: *age at diagnosis, Buffa Hypoxia Score, Cancer Type, Ethnicity Category, Neoplasm Histologic Grade, Overall Survival Status, Race Category, Radiation Therapy, Somatic Status, Tumour Type*, and *Patient Weight*. During the training phase, all the variables were utilized to predict the *overall survival status* (0: LIVING or 1: DECEASED). Table 1 shows the overall status; the dataset shape is 529; there are 115 missing values in the overall survival status; 359 of them relate to 0-LIVING and 55 of them relate to 1-DECESEAD.

TABLE I. Overall Survival Status

Feature	Details		
	Values	*0: Disease Free*	*1: Disease*
Overall Survival Status	414[a]	359	55

[a] Unique values

20% (106) of the data utilized in this study were for the testing phase, while the remaining data (423) were for the training phase.

B. Statistical Analysis

The statistical analysis was conducted using JASP (Version 0.17.1, 2023) in order to seek for trends, patterns, and correlations.

TABLE II. Descriptive Statistics

	Diagnosis Age	
	0:LIVING	1:DECEASED
Valid	439	87
Missing	3	0
Mean	63.171	66.770
Std. Deviation	10.987	11.003
Minimum	31.000	35.000
Maximum	90.000	90.000

According to Table II, diagnosis age, which depicts overall survival status, occurs in women who are around 66 years' age. But unfortunately, in the data, the earliest age is represented by a young adult who is 35 years old.

Moreover, a highly significant (*p-value <0.001*) relationship has been found between the age of diagnosis and the histologic grade of the tumor (G1, G2, G3, High Grade), race category (Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White), tumor type (Endometrioid Endometrial Adenocarcinoma, Mixed Serous and Endometrioid Carcinoma, Serous Endometrial Adenocarcinoma), subtype (UCEC_CN_LOW, UCEC_MSI, UCEC_POLE) according to Table III's ANOVA test.

TABLE III. ANOVA Test for Diagnosis Age

Cases	Sum of Squares	df	Mean Square	F	p
Neoplasm Histologic Grade	1834.24	3	611.41	5.11	0.002
Race Category	2916.89	4	729.22	6.48	< .001
Tumor Type	3688.45	2	1844.23	15.93	< .001
Subtype	7579.68	3	2526.56	23.41	< .001

Note. Type III Sum of Squares

Additionally, for all numeric variables, Pearson's correlation has been investigated, and the results can be observed in Table IV, which presents all variables that have a significant correlation (*p-value<0.05*). *The aneuploidy score* presented high significance related to features such as *diagnosis age*, *patient weight, mutation count, tumor mutational burden (TMB) (nonsynonymous), birth from the initial pathologic diagnosis date, fraction genome altered, MSI sensor score, and MSI MANTIS score*.

C. ML Model Performance

With the characteristics mentioned in the methods section as training data, the baseline models achieved 100% test accuracy.

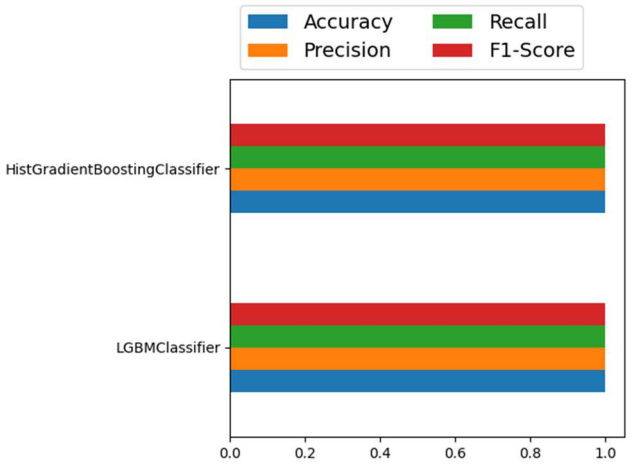

Fig. 3. ML Model performance.

To overcome the missing values in the dataset and achieve excellent performance, the performance metrics (accuracy, precision, recall, and F1-Score) for the two selected models—the *Histogram Gradient Boosting Classifier* and the *LGBM Classifier*—are shown in Fig. 3.

IV. Discussion

A key point in this dataset is represented by *aneuploidy score*. In almost common feature of human malignancies is aneuploidy, often known as whole chromosome or chromosomal arm imbalance. The expression of proliferative genes, somatic mutation rate, and TP53 mutation were shown to be linked with aneuploidy in 10,522 cancer genomes from The Cancer Genome Atlas (TCGA) [30-31]. Chromosome aneuploidy and driver mutations are two important factors that influence carcinogenesis and have intricate interactions [32].

[33] demonstrates a causal relationship between chemoresistance and alterations in gene copy number caused by aneuploidy, which may help to explain why some chemotherapies are ineffective. A connection has been shown lately between ageing, aneuploidy, and whole-chromosomal instability (W-CIN)[34]. Regardless of the TP53 genotype, individuals with aneuploid-high tumors had unfavorable prognoses, according to analysis of the outcome data. The findings suggest that the variety of previously documented mutant p53 GOF symptoms can be explained by genetic variation brought about by aneuploidy [35].

Nevertheless, the differential component is typically ignored by existing approaches [36], and traditional methods of differential analysis either cannot detect combinatorial patterns of difference or need a high data completeness rate. To do this, this study provides two strong and adaptable ML algorithms that may be used to analyze missing data.

TABLE IV. PEARSON'S CORRELATION

Feature	Feature	r	p
Aneuploidy Score	Diagnosis Age	0.230	< .001
Aneuploidy Score	Patient Weight	-0.102	0.024
Aneuploidy Score	Mutation Count	-0.229	< .001
Aneuploidy Score	TMB (nonsynonymous)	-0.229	< .001
Aneuploidy Score	Birth from Initial Pathologic Diagnosis Date	-0.229	< .001
Aneuploidy Score	Fraction Genome Altered	0.860	< .001
Aneuploidy Score	MSIsensor Score	-0.354	< .001
Aneuploidy Score	MSI MANTIS Score	-0.335	< .001
Diagnosis Age	Patient Weight	-0.252	< .001
Diagnosis Age	Mutation Count	-0.191	< .001
Diagnosis Age	TMB (nonsynonymous)	-0.196	< .001
Diagnosis Age	Birth from Initial Pathologic Diagnosis Date	-1.000	< .001
Diagnosis Age	Fraction Genome Altered	0.262	< .001
Patient Weight	Mutation Count	-0.176	< .001
Patient Weight	TMB (nonsynonymous)	-0.176	< .001
Patient Weight	Buffa Hypoxia Score	-0.225	0.005
Patient Weight	Progress Free Survival (Months)	-0.099	0.027
Patient Weight	Last Communication Contact from Initial Pathologic Diagnosis Date	-0.094	0.043
Patient Weight	Birth from Initial Pathologic Diagnosis Date	0.252	< .001
Patient Weight	Months of disease-specific survival	-0.096	0.031
Patient Weight	Overall Survival (Months)	-0.096	0.031
Patient Weight	Winter Hypoxia Score	-0.228	0.004
Mutation Count	TMB (nonsynonymous)	0.992	< .001
Mutation Count	Ragnum Hypoxia Score	0.244	0.001
Mutation Count	Progress Free Survival (Months)	0.254	< .001
Mutation Count	Last Communication Contact from Initial Pathologic Diagnosis Date	0.213	< .001
Mutation Count	Birth from Initial Pathologic Diagnosis Date	0.191	< .001
Mutation Count	Disease Free (Months)	0.250	< .001
Mutation Count	Months of disease-specific survival	0.209	< .001
Mutation Count	Overall Survival (Months)	0.209	< .001
Mutation Count	Fraction Genome Altered	-0.234	< .001
Mutation Count	MSIsensor Score	0.093	0.035
Mutation Count	MSI MANTIS Score	0.093	0.036
TMB (nonsynonymous)	Buffa Hypoxia Score	0.204	0.008
TMB (nonsynonymous)	Ragnum Hypoxia Score	0.240	0.002
TMB (nonsynonymous)	Progress Free Survival (Months)	0.252	< .001
TMB (nonsynonymous)	Last Communication Contact from Initial Pathologic Diagnosis Date	0.210	< .001
TMB (nonsynonymous)	Birth from Initial Pathologic Diagnosis Date	0.196	< .001
TMB (nonsynonymous)	Disease Free (Months)	0.246	< .001
TMB (nonsynonymous)	Months of disease-specific survival	0.206	< .001
TMB (nonsynonymous)	Overall Survival (Months)	0.206	< .001
TMB (nonsynonymous)	Fraction Genome Altered	-0.234	< .001
TMB (nonsynonymous)	MSIsensor Score	0.089	0.043
TMB (nonsynonymous)	MSI MANTIS Score	0.088	0.045
Buffa Hypoxia Score	Ragnum Hypoxia Score	0.724	< .001
Buffa Hypoxia Score	Winter Hypoxia Score	0.823	< .001
Buffa Hypoxia Score	MSI MANTIS Score	0.157	0.038
Ragnum Hypoxia Score	Winter Hypoxia Score	0.666	< .001
Ragnum Hypoxia Score	MSI MANTIS Score	0.164	0.031
Progress Free Survival (Months)	Last Communication Contact from Initial Pathologic Diagnosis Date	0.892	< .001
Progress Free Survival (Months)	Disease Free (Months)	0.990	< .001
Progress Free Survival (Months)	Months of disease-specific survival	0.909	< .001
Progress Free Survival (Months)	Overall Survival (Months)	0.909	< .001
Progress Free Survival (Months)	Fraction Genome Altered	-0.121	0.006
Last Communication Contact from Initial Pathologic Diagnosis Date	Disease Free (Months)	0.900	< .001
Last Communication Contact from Initial Pathologic Diagnosis Date	Months of disease-specific survival	0.973	< .001

Last Communication Contact from Initial Pathologic Diagnosis Date	Overall Survival (Months)	0.973	< .001
Last Communication Contact from Initial Pathologic Diagnosis Date	Fraction Genome Altered	-0.107	0.019
Birth from Initial Pathologic Diagnosis Date	Fraction Genome Altered	-0.262	< .001
Disease Free (Months)	Months of disease-specific survival	0.915	< .001
Disease Free (Months)	Overall Survival (Months)	0.915	< .001
Months of disease-specific survival	Overall Survival (Months)	1.000	< .001
Months of disease-specific survival	Fraction Genome Altered	-0.106	0.016
Overall Survival (Months)	Fraction Genome Altered	-0.106	0.016
Fraction Genome Altered	MSIsensor Score	-0.352	< .001
Fraction Genome Altered	MSI MANTIS Score	-0.332	< .001
Winter Hypoxia Score	MSIsensor Score	0.180	0.017
MSIsensor Score	MSI MANTIS Score	0.948	< .001

The *Buffa Hypoxia Score, Ragnum Hypoxia Score, MSI MANTIS Score and MSI sensor score, Progress Free Survival (Months), Last Communication Contact from Initial Pathologic Diagnosis Date, Birth from Initial Pathologic Diagnosis Date, Disease Free (Months), Months of disease-specific survival, and Overall Survival (Months)* all showed that TMB was highly significant (Table 4). It has been suggested that TMB may serve as a predictive biomarker for immunotherapy in a variety of solid tumors, based on the growing data, although more prospective validation needs to be conducted. The usual practice of oncology now includes the use of TMB [37].

V. Conclusions and Future work

An important contributor to the morbidity and death associated with cancer is EC, a frequent uterine malignancy. Treatment response is low for EC found at an advanced stage.

Large-scale data analysis, personalized medical strategies, and multidisciplinary cooperation are some of the ways that ML techniques have the potential to transform endometrial cancer diagnosis, treatment, and prognosis. Ultimately, the aim is to lessen the impact that endometrial cancer has on both individuals and society by utilizing ML to enhance patient outcomes and augment scientific understanding.

ML approaches might be critical to the diagnosis, treatment, and prediction of endometrial carcinoma. Prior to delving into an exhaustive genetic study of viable ML approaches for EC prevention, screening, detection, and prognosis, this work provides an overview of EC, risk factors, and diagnostic tools by utilizing statistical analysis and ML across a complex dataset.

In order to address missing values in the data that might affect performance and produce a more robust solution, we would like to carry out in-depth assessments of various potential algorithms as a future research topic.

References

[1] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA. Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, 2024, doi: 10.3322/caac.21820.

[2] "Centrul de Calcul, Statistică Sanitară şi Documentare Medicală: Registrul Naţional de Cancer, MSP, Bucureşti," 2004.

[3] J. A. Lachance, C. J. Darus, and L. W. Rice, "Surgical management and postoperative treatment of endometrial carcinoma," *Rev. Obstet. Gynecol.*, vol. 1, no. 3, pp. 97–105, 2008.

[4] D. E. Telner and D. Jakubovicz, "Approach to diagnosis and management of abnormal uterine bleeding," *Can. Fam. Physician Med. Fam. Can.*, vol. 53, no. 1, pp. 58–64, Jan. 2007.

[5] J. A. Nyaaba and E. Akurugu, "Knowledge, barriers and uptake towards Cervical Cancer screening among female health workers in Ghana: A perspective of the Health Belief Model," *Int. J. Afr. Nurs. Sci.*, vol. 19, p. 100587, 2023, doi: 10.1016/j.ijans.2023.100587.

[6] R. Navaratnarajah, O. Pillay, and P. Hardiman, "Polycystic Ovary Syndrome and Endometrial Cancer," *Semin. Reprod. Med.*, vol. 26, no. 1, pp. 062–071, Jan. 2008, doi: 10.1055/s-2007-992926.

[7] V. W. Setiawan *et al.*, "Age at Last Birth in Relation to Risk of Endometrial Cancer: Pooled Analysis in the Epidemiology of Endometrial Cancer Consortium," *Am. J. Epidemiol.*, vol. 176, no. 4, pp. 269–278, Aug. 2012, doi: 10.1093/aje/kws129.

[8] M. Lambe, J. Wuu, E. Weiderpass, and C.-C. Hsieh, "Childbearing at older age and endometrial cancer risk (Sweden)," *Cancer Causes Control*, vol. 10, no. 1, pp. 43–49, 1999, doi: 10.1023/A:1008860615584.

[9] R. M. Pfeiffer *et al.*, "Timing of births and endometrial cancer risk in Swedish women," *Cancer Causes Control*, vol. 20, no. 8, pp. 1441–1449, Oct. 2009, doi: 10.1007/s10552-009-9370-7.

[10] M. Hinkula, E. Pukkala, P. Kyyrönen, and A. Kauppila, "Grand multiparity and incidence of endometrial cancer: A population-based study in Finland," *Int. J. Cancer*, vol. 98, no. 6, pp. 912–915, Apr. 2002, doi: 10.1002/ijc.10267.

[11] A. T. Ali, "Reproductive Factors and the Risk of Endometrial Cancer," *Int. J. Gynecol. Cancer*, vol. 24, no. 3, pp. 384–393, Mar. 2014, doi: 10.1097/IGC.0000000000000075.

[12] C. P. McPherson, T. A. Sellers, J. D. Potter, R. M. Bostick, and A. R. Folsom, "Reproductive Factors and Risk of Endometrial Cancer The Iowa Women's Health Study," *Am. J. Epidemiol.*, vol. 143, no. 12, pp. 1195–1202, Jun. 1996, doi: 10.1093/oxfordjournals.aje.a008707.

[13] S. J. Jordan *et al.*, "Breastfeeding and Endometrial Cancer Risk: An Analysis From the Epidemiology of Endometrial Cancer Consortium," *Obstet. Gynecol.*, vol. 129, no. 6, pp. 1059–1067, Jun. 2017, doi: 10.1097/AOG.0000000000002057.

[14] X. Ma *et al.*, "Association between breastfeeding and risk of endometrial cancer: a meta-analysis of epidemiological studies," *Eur. J. Cancer Prev.*, vol. 27, no. 2, pp. 144–151, Mar. 2018, doi: 10.1097/CEJ.0000000000000186.

[15] M. Crous-Bou *et al.*, "Coffee consumption and risk of endometrial cancer: a pooled analysis of individual participant data in the Epidemiology of Endometrial Cancer Consortium (E2C2)," *Am. J. Clin. Nutr.*, vol. 116, no. 5, pp. 1219–1228, Nov. 2022, doi: 10.1093/ajcn/nqac229.

[16] L. Costas *et al.*, "Night work, chronotype and risk of endometrial cancer in the Screenwide case–control study," *Occup. Environ. Med.*, vol. 79, no. 9, pp. 624–627, Sep. 2022, doi: 10.1136/oemed-2021-108080.

[17] L. Zheng *et al.*, "Latest advances in the study of non-coding RNA-mediated circadian rhythm disorders causing endometrial cancer," *Front. Oncol.*, vol. 13, p. 1277543, Nov. 2023, doi: 10.3389/fonc.2023.1277543.

[18] M. W. Greene, "Circadian rhythms and tumor growth," *Cancer Lett.*, vol. 318, no. 2, pp. 115–123, May 2012, doi: 10.1016/j.canlet.2012.01.001.

[19] M. C. Playdon *et al.*, "Alcohol and oestrogen metabolites in postmenopausal women in the Women's Health Initiative

Observational Study," *Br. J. Cancer*, vol. 118, no. 3, pp. 448–457, Feb. 2018, doi: 10.1038/bjc.2017.419.

[20] A. Lafranconi *et al.*, "Coffee Decreases the Risk of Endometrial Cancer: A Dose–Response Meta-Analysis of Prospective Cohort Studies," *Nutrients*, vol. 9, no. 11, p. 1223, Nov. 2017, doi: 10.3390/nu9111223.

[21] M. Hashibe *et al.*, "Coffee, tea, caffeine intake, and the risk of cancer in the PLCO cohort," *Br. J. Cancer*, vol. 113, no. 5, pp. 809–816, Sep. 2015, doi: 10.1038/bjc.2015.276.

[22] S. J. Kitson, D. G. Evans, and E. J. Crosbie, "Identifying High-Risk Women for Endometrial Cancer Prevention Strategies: Proposal of an Endometrial Cancer Risk Prediction Model," *Cancer Prev. Res. (Phila. Pa.)*, vol. 10, no. 1, pp. 1–13, Jan. 2017, doi: 10.1158/1940-6207.CAPR-16-0224.

[23] A. A. Secord *et al.*, "Body mass index and mortality in endometrial cancer: A systematic review and meta-analysis," *Gynecol. Oncol.*, vol. 140, no. 1, pp. 184–190, Jan. 2016, doi: 10.1016/j.ygyno.2015.10.020.

[24] "Mahdy H, Casey MJ, Crotzer D. Endometrial Cancer. [Updated 2022 Sep 26]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK525981/."

[25] V. Makker *et al.*, "Endometrial cancer," *Nat. Rev. Dis. Primer*, vol. 7, no. 1, p. 88, Dec. 2021, doi: 10.1038/s41572-021-00324-8.

[26] The Cancer Genome Atlas Research Network and D. A. Levine, "Integrated genomic characterization of endometrial carcinoma," *Nature*, vol. 497, no. 7447, pp. 67–73, May 2013, doi: 10.1038/nature12113.

[27] V. Bruno *et al.*, "Machine learning endometrial cancer risk prediction model: integrating guidelines of European Society for Medical Oncology with the tumor immune framework," *Int. J. Gynecol. Cancer*, vol. 33, no. 11, pp. 1708–1714, Nov. 2023, doi: 10.1136/ijgc-2023-004671.

[28] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[29] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[30] A. M. Taylor *et al.*, "Genomic and Functional Approaches to Understanding Cancer Aneuploidy," *Cancer Cell*, vol. 33, no. 4, pp. 676-689.e3, Apr. 2018, doi: 10.1016/j.ccell.2018.03.007.

[31] L.-F. Herlo *et al.*, "Colorectal Cancer Risk Prediction Using the rs4939827 Polymorphism of the SMAD7 Gene in the Romanian Population," *Diagnostics*, vol. 14, no. 2, p. 220, Jan. 2024, doi: 10.3390/diagnostics14020220.

[32] N. Auslander, Y. I. Wolf, and E. V. Koonin, "Interplay between DNA damage repair and apoptosis shapes cancer evolution through aneuploidy and microsatellite instability," *Nat. Commun.*, vol. 11, no. 1, p. 1234, Mar. 2020, doi: 10.1038/s41467-020-15094-2.

[33] M. R. Ippolito *et al.*, "Gene copy-number changes and chromosomal instability induced by aneuploidy confer resistance to chemotherapy," *Dev. Cell*, vol. 56, no. 17, pp. 2440-2454.e6, Sep. 2021, doi: 10.1016/j.devcel.2021.07.006.

[34] R. M. Naylor and J. M. Van Deursen, "Aneuploidy in Cancer and Aging," *Annu. Rev. Genet.*, vol. 50, no. 1, pp. 45–66, Nov. 2016, doi: 10.1146/annurev-genet-120215-035303.

[35] L. N. Redman-Rivera *et al.*, "Acquisition of aneuploidy drives mutant p53-associated gain-of-function phenotypes," *Nat. Commun.*, vol. 12, no. 1, p. 5184, Aug. 2021, doi: 10.1038/s41467-021-25359-z.

[36] L. Zhang and S. Zhang, "Learning common and specific patterns from data of multiple interrelated biological scenarios with matrix factorization," *Nucleic Acids Res.*, vol. 47, no. 13, pp. 6606–6617, Jul. 2019, doi: 10.1093/nar/gkz488.

[37] D. Sha, Z. Jin, J. Budczies, K. Kluck, A. Stenzinger, and F. A. Sinicrope, "Tumor Mutational Burden as a Predictive Biomarker in Solid Tumors," *Cancer Discov.*, vol. 10, no. 12, pp. 1808–1825, Dec. 2020, doi: 10.1158/2159-8290.CD-20-0522.