healthcare

MDPI

Article

A Noninvasive Risk Stratification Tool Build Using an Artificial Intelligence Approach for Colorectal Polyps Based on Annual Checkup Data

Chieh Lee [1], Tsung-Hsing Lin [2], Chen-Ju Lin [3], Chang-Fu Kuo [4], Betty Chien-Jung Pai [5], Hao-Tsai Cheng [6], Cheng-Chou Lai [7] and Tsung-Hsing Chen [8,]*

[1] Department of Information Management, National Sun Yat-sen University, Kaohsiung 804, Taiwan; chiehlee850427@gmail.com
[2] Department of Emergency Medicine, Kuang Tien General Hospital, Taichung City 433, Taiwan; drsixmg@gmail.com
[3] Department of Industrial Engineering & Management, College of Engineering, Yuan Ze University, Chung-Li City 320, Taiwan; chenju.lin@saturn.yzu.edu.tw
[4] Division of Rheumatology, Allergy, and Immunology, Linkou Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Taoyuan 333, Taiwan; zandis@gmail.com
[5] Craniofacial Orthodontics, Craniofacial Research Center, Chang Gung Memorial Hospital, Chang Gung University, Taoyuan 333, Taiwan; pai0072@cgmh.org.tw
[6] Division of Gastroenterology and Hepatology, Department of Internal Medicine, New Taipei Municipal TuCheng Hospital, New Taipei City 236, Taiwan; hautai@cloud.cgmh.org.tw
[7] Department of Colon and Rectal Surgery, Linkou Medical Center, Chang Gung Memorial Hospital, Taoyuan 333, Taiwan; lai5556@cgmh.org.tw
[8] Department of Gastroenterology and Hepatology, Linkou Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Taoyuan 333, Taiwan
* Correspondence: itochenyu@gmail.com

check for updates

Citation: Lee, C.; Lin, T.-H.; Lin, C.-J.; Kuo, C.-F.; Pai, B.C.-J.; Cheng, H.-T.; Lai, C.-C.; Chen, T.-H. A Noninvasive Risk Stratification Tool Build Using an Artificial Intelligence Approach for Colorectal Polyps Based on Annual Checkup Data. *Healthcare* **2022**, *10*, 169. https://doi.org/10.3390/healthcare10010169

Academic Editors: Mahmudur Rahman and Francesco Faita

Received: 10 November 2021
Accepted: 12 January 2022
Published: 17 January 2022

Abstract: Colorectal cancer is the leading cause of cancer-related deaths worldwide, and early detection has proven to be an effective method for reducing mortality. The machine learning method can be implemented to build a noninvasive stratifying tool that helps identify patients with potential colorectal precancerous lesions (polyps). This study aimed to develop a noninvasive risk-stratified tool for colorectal polyps in asymptomatic, healthy participants. A total of 20,129 consecutive asymptomatic patients who underwent a health checkup between January 2005 and August 2007 were recruited. Positive relationships between noninvasive risk factors, such as age, *Helicobacter pylori* infection, hypertension, gallbladder polyps/stone, and BMI and colorectal polyps were observed ($p < 0.0001$), regardless of sex, whereas significant findings were noted in men with tooth disease ($p = 0.0053$). A risk stratification tool was developed, for colorectal polyps, that considers annual checkup results from noninvasive examinations. For the noninvasive stratified tool, the area under the receiver operating characteristic curve (AUC) of obese females (males) aged <50 years was 91% (83%). In elderly patients (>50 years old), the AUCs of the stratifying tools were >85%. Our results indicate that the risk stratification tool can be built by using random forest and serve as an efficient noninvasive tool to identify patients requiring colonoscopy.

Keywords: *Helicobacter pylori* infection; colorectal polyp; teeth disease; precancerous lesions; noninvasive; risk stratifying tool; random forest

1. Introduction

Colorectal cancer (CRC) is the most common cancer worldwide and a significant public health problem in developed countries [1,2]. Most CRCs arise from polyps considered to be precancerous lesions, particularly adenomatous polyps [3–6], even though most are asymptomatic. Removal of all precancerous lesions during endoscopy has been the most effective method for preventing cancer development [6–8]. Colonoscopy is the most effective method for the search and removal of colorectal polyps. However, colonoscopy is not only time consuming and costly but also has side effects. Previous studies have reported several adverse events of colonoscopy, including perforation (0.005–0.085%) and bleeding (0.0001–0.687%) [9]. These adverse events create health hazards for patients and financial burdens for healthcare centers.

Furthermore, the increasing demand for colonoscopy drastically increases the workload of gastroenterology [10]. The increasing workload might result in undesired results such as lower adenoma detection rates per colonoscopy [11] and longer waiting times for colonoscopy [12]. As shown in [12], the median waiting time for the screening colonoscopy is 210 days with the maximum waiting time equaling 631 days in Canada. Long waiting times increases the patient's mental burden and the risk of precancerous polyps' evolvement. Therefore, healthcare centers are actively searching for a risk stratification tool that identifies patients who require colonoscopy using noninvasive examination results.

Hence, risk factors of noninvasive examination data for colorectal polyps, such as gender, age, BMI, blood pressure, gallbladder (GB) polyp/stone, *Helicobacter pylori* infection, and tooth disease (periodontal disease, chronic gingivitis, and chronic periodontitis), were collected, and a machine learning method was implemented to build a risk stratification tool for patients with colorectal polyps. Risk factors were selected based on previous studies [13–17], which reported factors exhibiting some relationship with precancerous polyps [18]. Data from 20,129 consecutive asymptomatic individuals who underwent a health checkup were collected. To date, little is known about their association. Here, we hypothesized that noninvasive risk factors may be associated with colorectal precancerous lesions. Furthermore, we hypothesized that risk factors might vary from patients groups with different demographic characteristics such as gender, age, weights, etc.

After identifying noninvasive risk factors and patient grouping criteria, a noninvasive risk stratification tool was built in order to identify patients who need colonoscopy using a machine learning method. Previous studies have investigated the possibility of identifying patients at high risk for heart disease [19] and diabetes [20] using machine learning methods. More recently, artificial intelligence approaches such as machine learning methods have been used to build a risk stratification tool for different diseases [21]. Therefore, based on the identified risk factors, a machine learning method was further employed to show that the identified risk factors can serve as predictors of precancerous lesions.

To the best of our knowledge, this is the first investigation aimed at building a noninvasive stratification tool based on risk factors from annual checkup data. This study aimed to develop a simple, noninvasive, risk factor, and noninvasive risk stratification tool for these asymptomatic populations to determine colorectal precancerous lesions.

2. Materials and Methods

2.1. Study Participants

In this retrospective study, 20,129 consecutive asymptomatic patients who underwent a health checkup between January 2005 and August 2007 at Chang Gung Memorial Hospital (approval number: 201601348B0, approved 2016/01) were recruited. This study was approved by the Ethics Committee of the Institutional Review Board of Chang Gung Memorial Hospital and conducted according to the ethical principles of the Declaration of Helsinki, as reflected in the a priori approval by the institution's human research committee. Written informed consent was obtained from all patients included in the study. Our health checkup program included physical examination, chest radiography, electrocardiography, complete blood tests, biochemical laboratory tests, urine analysis, abdominal ultrasonography, and colonoscopy. Exclusion criteria were patients who did not have colonoscopy during the course of the health checkup or had incomplete colonoscopy due to various reasons, such as poor bowel preparation or incomplete total colon inspection and BMI > 35 kg/m^2. Height and body weight, used to calculate BMI, were measured by well-trained nurses. BMI ranges were underweight, under 18.5 kg/m^2; normal weight, 18.5–25 kg/m^2; overweight, 25–30 kg/m^2; and obese, >30 kg/m^2. In our institution, the C13 urea breath test was used to detect *Helicobacter pylori* infection [22].

2.2. Colonoscopy Procedure and Abdominal Ultrasonography

For bowel preparation, patients ingested 1.5–2 L of polyethylene glycol before the procedure. All procedures were performed by experienced gastroenterologists. Endoscopic findings were classified into two subgroups: polyp and polyp-free. GB polyps on ultrasonography showed fixed, hyperechoic material attached to the lumen of the GB, without an acoustic shadow [23].

2.3. Risk Stratification Tool Building

As described in Section 2.1, all items in the annual check-up data are collected for this research. Based on previous research [13–17], we selected risk factors from the following categories: (1) patient's demographic characteristics including age, sex, weight, and height; (2) patient's medical history including hypertension, diabetes, and *Helicobacter pylori* infection; (3) colonoscopy diagnosis results including colorectal polyps, ulcerative colitis, hemorrhoids, and intestinal hemorrhage, etc.; (4) abdominal ultrasonography diagnosis including GB polyps and GB stones; (5) blood sample diagnosis results including fasting blood glucose, total cholesterol, high and low-density lipoprotein (HDL and LDL), triglycerides, etc.; (6) dental diagnosis results including periodontitis, periodontal disease, chronic periodontitis, and chronic gingivitis. All diagnosis results are binary with respect to data with 1 = positive diagnosed and 0 = otherwise. BMI is calculated based on the weight of height of the patient. Furthermore, patients' demographic data are dichotomized into binary or categorical data. Age is dichotomized as over (1)/under (0) 50 years old, and BMI is categorized as 0 (underweight (<18.5 kg/m^2)), 1 (normal (18.5–25 kg/m^2)), 2 (overweight (25–30 kg/m^2)), and 3 (obese (>30 kg/m^2)).

Our overall risk stratification tool building procedure is summarized in Figure 1 and *the Heuristic*.

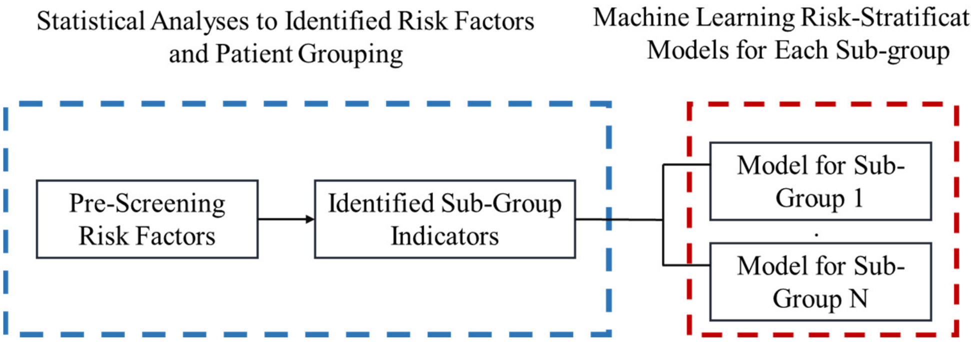
Statistical Analyses to Identified Risk Factors and Patient Grouping
Machine Learning Risk-Stratification Models for Each Sub-group
Pre-Screening Risk Factors
Identified Sub-Group Indicators
Model for Sub-Group 1
Model for Sub-Group N

Figure 1. Diagram for proposed *Heuristic*.

The Heuristic:

Step 1: Collect data from annual health check-ups. All risk factors are indexed from $i = 1 \ldots N$, the value of the risk factor is x_i, where there are N risk factors in total.

Step 2: Pre-screen with a z-test for two sample proportions with a significance level equal to 0.05 is applied to select potential risk factors. Where the two sample proportions are calculated as For all risk factor i,

p_{hi} = the proportion of patients who has colorectal polyps for patients with risk factor $x_i = h - 1$.

That is,

p_{1i} = the proportion of patients who has colorectal polyps for patients with risk factor $x_i = 0$.

p_{2i} = the proportion of patients who has colorectal polyps for patients with risk factor $x_i = 1$.

Step 3: The null and alternative hypothesis is stated as below: Null Hypothesis: $p_{1i} = p_{2i} = \dots . p_{\mathrm{h}i}$

We record all risk factors which has a significantly different sample proportion between patients with and without colorectal polyps.

Step 4: Logistics regression is applied for each risk factor to calculate the discriminability for each risk factor. Based on the logistic regression, we identified the demographic risk factors which can segregate patients into different sub-groups for the machine learning process.

Step 5: Machine learning is applied to each sub-group to construct the risk stratification tool.

Step 6: We output the system of models which consisted of multiple random forest models.

Step 7: Output our four-fold-cross validation.

2.4. Statistical Analyses

Statistical analyses, including receiver operating characteristic (ROC) curve, area under ROC (AUC), multinomial logistic regression analyses, and z-test for two-sample proportions, were conducted using SAS software (version 9.4; SAS Institute, Cary, NC, USA). We use the two-sample z-test for the pre-screen tool since it is simple and efficient. Researchers might consider another pre-screen method as well. Statistical significance was set at $p < 0.05$. Simple logistic regression was applied when the independent risk factor was binary (e.g., age), and multinomial logistic regression was applied when the independent risk factor was categorical (e.g., BMI). The AUC was reported for each logistic regression. Since underweight, overweight, and obesity groups were all considered abnormal, BMI was treated as categorical instead of ordinal data. Tooth disease was identified if the patient was diagnosed with periodontal disease, chronic periodontal disease, and/or chronic gingivitis. GB equaled a score of one if GB polyps and stones were observed on abdominal ultrasonography, whereas hypertension was based on the patient's medical history and not the onsite measurement of blood pressure.

2.5. Machine Learning Algorithm

A machine learning algorithm, random forest, was adopted by using Python to build a risk stratification tool based on the risk factors identified from annual healthcare data. Discriminability was represented by AUCs. We used 75% of the data to build the model and 25% of the data to test the consistency of the model. The model building and testing process was repeated four times (four-fold validation method). Adulqader et al. [14] conducted a review on machine learning in healthcare. The authors point out the most popular classification method among all machine learning algorithms including support vector machine (SVM), random forest (RF), and Naïve Bayes. Previous studies [24–26] also use annual health check-up data to develop a risk stratification tool to serve as a screening tool for non-alcoholic fatty liver disease. Goldman et al. [25] use the decision-tree-based approach, and Fialoke et al. [26] used several other methods along with the decision-tree approach. We argue that since our risk factors are all binary data, a decision tree-based method such as RF is the most suitable method. Our machine learning algorithm is summarized as the following pseudo-code.

Machine Learning Algorithm (RF):

Step 1: Input all risk factors as vector X = <x1 . . . . . . xh> and the y = 1 if a patient is diagnosed with colorectal polyps, and zero otherwise. Moreover, input the demographic factors for aggregating patients into subgroups. Go to Step 2.

Step 2: Segregate all patients into subgroups. Index subgroups as k = 1 . . . N for N groups in total. Let k =1 and go to Step 3.

Step 3: Input all risk factors X and y in the kth sub-group. Go to Step 4.

Step 4: Input all data in with path_name = group k, with the following specification of random foreackage in python. We selected the four-fold validation, thus 75% of data will be randomly selected for modeling building and 25% will be reserved for validation. For each run, the random forest will repeat four times for validation. Output the model and go to Step 5. Branch criterion: gini index Number of estimators (number of decision trees): 1000 Min_samples_leaf = 5 Class weight: balanced Validation: Four-fold Calculate the following statistics: Specificity = True negative/(true negative + false positive) Sensitivity = True positive/(true positive + false negative) Area Under Curve (AUC)

Step 5: Collected the outputted model and check if k = N, if not let k = k + 1 and go to Step 3, otherwise end the algorithm.

It is worth noting that all parameters are subjected to test and modified for different research topics. The parameters provided in the algorithm are the optimal parameters after our testing trials.

3. Results

3.1. Statistical Analysis

A total of 20,129 patients were enrolled, including 11,570 (57.5%) men and 8559 (42.5%) women, with a median age of 50 (range: 18–96) years, GB polyps/stones (3191, 15.85%), and tooth disease (15,346, 76.24%), as shown in Table 1. In this study, the risk factors of colorectal polyps were investigated. Each group was subdivided into two groups based on endoscopic findings: polyp and polyp-free. Logistic regression analysis was performed after adjusting for age, gender, BMI, GB polyp/stone, tooth disease, hypertension, and *Helicobacter pylori* infection to determine independent predictors of colorectal polyps. The prevalence of colorectal polyps was 27.08% (5450/20,129) and was associated with age, *Helicobacter pylori* infection, hypertension, and BMI (underweight and overweight) regardless of sex ($p < 0.0001$). Tooth disease only showed a significant difference in men ($p = 0.0053$), as shown in Table 2.

Table 1. Participants' clinical characteristics.

Total Number	***n*, %**	**20,129**
Gender	Ratio of male to female (*n*/*n*)	11,570:8559
Polyp		
	Colorectal polyp (*n*, %)	5450, 27.08%
	Gallbladder polyps (*n*, %)	2188, 10.87%
	Gallbladder stone (*n*, %)	1106, 5.49%
Gallbladder problem		3191, 15.85%
Hypertension	(*n*, %)	1684, 8.37%
Helicobacter pylori infection	(*n*, %)	751, 3.73%
Tooth disease		15,346, 76.24%
	Periodontal disease (*n*, %)	8917, 44.30%
	Chronic gingivitis (*n*, %)	4168, 20.71%
	Chronic periodontitis (*n*, %)	11,655, 57.90%
BMI		
	Underweight (*n*, %)	805, 4%
	Normal (*n*, %)	9090, 45.16%
	Overweight (*n*, %)	6046, 30.04%
	Obesity (*n*, %)	4188, 20.81%
Age	Median (range)	50 (18–96) years
Total cholesterol		2818, 14%
HDL		2617, 13%
Triglycerides		3452, 17%

Table 2. Multinomial logistic regression analysis of variables for colorectal polyps.

		Regardless of Gender		**Male**		**Female**	
Parameters		***p*-Value**	**AUC**	***p*-Value**	**AUC**	***p*-Value**	**AUC**
Age	(>50 years = 1)	<0.0001	0.5847	<0.0001	0.5906	<0.0001	0.5900
Helicobacter pylori	(Yes = 1)	<0.0001	0.5113	<0.0001	0.5104	<0.0001	0.5092
Hypertension	(Yes = 1)	<0.0001	0.5142	0.0029	0.5084	<0.0001	0.5240
Tooth disease	Total	0.3734	0.503	0.0053	0.5118	0.1041	0.5086
Gallbladder	(Yes = 1)	<0.0001	0.514	0.002	0.5119	0.0185	0.5105
BMI							
	Underweight = 0	<0.0001	0.5604	0.0012	0.5389	<0.0001	0.5709
	Normal = 1	0.0055		0.1301		0.0341	
	Overweight = 2	<0.0001		0.0017		0.008	
	Obesity = 3						

In Table 2, we find that the risk factors differ based on gender, age, and BMI. Therefore, all patients were divided into sub-groups based on gender, age, and BMI. For each group, risk factors for GB polyps, hypertension, tooth, disease, and *Helicobacter pylori* infection were input as independent variables to predict colorectal polyps. In Table 2 we presented the AUC of risk factors with *p*-values of the model and AUC from the logistics equations, where the *p*-values are less than 0.1 for at least male or female. Results of total cholesterol, high lipoprotein cholesterol, and triglycerides are excluded since their *p*-values are greater than 0.1. As we can observe from Table 2, the observed significances (*p*-values) for risk factors are different from male to female. Thus, we separate patients with their gender for the machine learning stage. While in Table 2 we did not examine the *p*-value for different BMI levels, previous literature suggests BMI might significantly relate to the evolvement of colorectal polyps. For example, [27] found that overweight and underweight statuses are significantly correlated with gut microbiota and metabolism. Jain et al. [28] found that obesity significantly impacts metabolism and is accessible with colorectal cancer and polyps. Hence, we also separate patients with their status of BMI.

Figures 2 and 3 further demonstrate the significance and positive or negative impacts of each risk factor, respectively. In order to construct a risk stratification tool based on these risk factors, a random forest machine learning method was employed. In our study, age, *Helicobacter pylori* infection, and hypertension were all risk factors for colorectal polyps. A forest chart was also constructed to present estimated odds ratios for each risk factor, as shown in Figures 2 and 3. While traditional statistical methods such as logistic regression have an AUC > 0.5, discriminability is not as high as healthcare centers may wish (0.5086–0.5900). Therefore, a machine learning method is required to build a model with higher discriminability. As shown in Figures 2 and 3, abnormal body mass, age, and *Helicobacter pylori* are the most influential risk factors for colorectal polyps, regardless of the patient's gender. We also found that hypertension was a significant risk factor for colorectal polyps in male patients. Moreover, the influence of different abnormal body masses was significantly different between gender and age groups. Thus, we further divided patients according to age, gender, and body mass to obtain 16 patient subgroups (2 × 2 × 4). Since risk factors differ according to age and sex, a risk stratification model was built for each group of patients. For each subgroup, a risk stratification tool was built via a machine learning method. Building a patient-characteristic-specific risk stratification model by using the machine learning method not only enhances the discriminability of the model but also identifies a set of more precise risk factors for each patient group. Healthcare centers can utilize these risk factors to precisely diagnose patients with colorectal polyps.

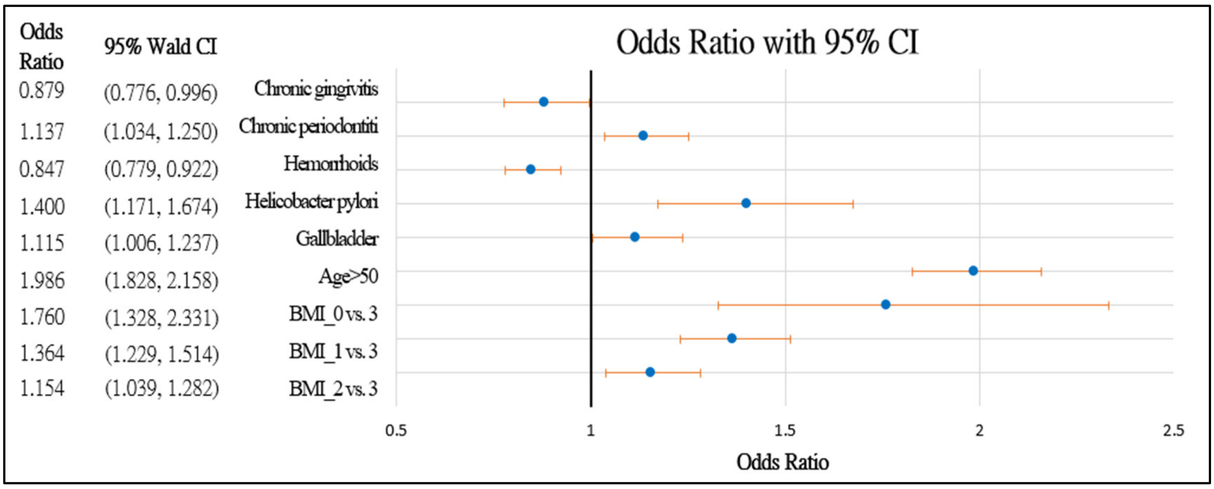

Figure 2. Forest chart of colorectal polyps' risk factors in female patients. Underweight = 0, normal = 1, overweight = 2, and obesity = 3.

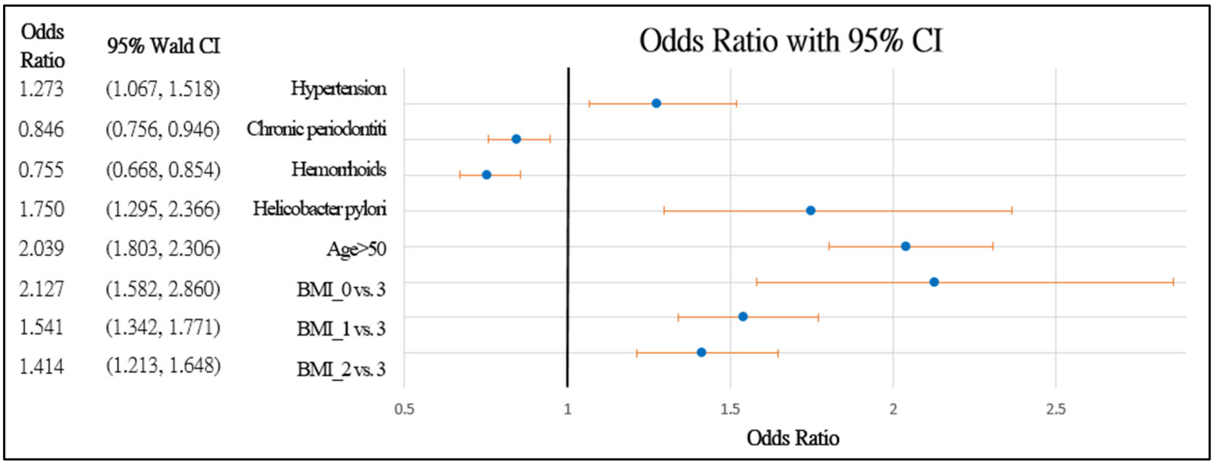

Figure 3. Forest chart of colorectal polyps' risk factors in male patients. Underweight = 0, normal = 1, overweight = 2, and obesity = 3.

3.2. Noninvasive Diagnostics Tool with Random Forests

Based on our results in Section 3.1, we separate all patients into 16 groups via their age, gender, and BMI status. The random forest algorithm in Section 2.5 is applied to each group, and validation results are summarized in Table 3. The input risk factors include hypertension, chronic periodontitis, humanoids, *Helicobacter pylori* infection, GB stones and polyps, total cholesterol, high-density lipoprotein, triglycerides, and diabetes. However, not all risk factors are significant in the final model, and the performance of the stratification model varied extensively. In women < 50 years old with a BMI > 30 kg/m^2, the random forest model's discriminability (AUC = 91%) was high compared to that in other groups. The discriminability of detecting colorectal polyps is >80% for both women and men who are obese. The noninvasive detection tool has an AUC = 80% for underweight male who is >50 years old. In general, the noninvasive colorectal polyp detection tool has a higher AUC in patients with abnormal weight.

Table 3. Noninvasive stratifying tool (random forests model).

Gender	Age	BMI	Sensitivity	Specificity	AUC
Female	<50 years old	Normal	0.22	**0.74**	0.61
		Overweight	0.09	**0.83**	**0.76**
		Obese	0.14	**0.79**	**0.91**
		Underweight	0.55	0.50	0.66
	≥50 years old	Normal	0.35	0.66	0.68
		Overweight	0.27	**0.74**	0.68
		Obese	0.34	**0.74**	**0.85**
		Underweight	0.05	0.67	**0.79**
Male	<50 years old	Normal	0.38	0.68	0.63
		Overweight	0.39	0.59	0.68
		Obese	0.29	0.67	**0.83**
		Underweight	0.11	**0.72**	**0.75**
	≥50 years old	Normal	0.56	0.47	0.67
		Overweight	0.47	0.52	**0.70**
		Obese	0.43	0.57	**0.87**
		Underweight	0.28	0.65	**0.80**

Furthermore, important risk factors identified by the random forests were examined. As shown in Table 3, in women aged >50 years and BMI > 18.5 kg/m^2, the important risk factors are hypertension, diabetes, and GB stones. In contrast, in women <50 years of age and BMI >18.5 kg/m^2, the important risk factors are GB stones and polyps. In men, for those >50 years of age and not underweight, the important risk factors are hypertension, diabetes, and high-density cholesterol. In men aged <50 years, the important risk factors are total cholesterol and high-density cholesterol. As observed, GB polyps and stones are important risk factors for predicting colorectal polyps in female patients.

4. Discussion

To the best of our knowledge, this is the first retrospective study to construct a noninvasive stratification tool for colorectal polyps based on an extensive set of risk factors identified by evaluating a possible association between colorectal polyps, GB polyps/stone, and tooth disease in healthy individuals. In this study, the participants were divided into two groups: polyp and polyp-free. Age, gender, BMI, GB polyps/stone, tooth disease (periodontal disease, chronic gingivitis, and chronic periodontitis), colorectal polyp, hypertension, and *Helicobacter pylori* infection; and triglyceride, high-density lipoprotein cholesterol, and total cholesterol were investigated. Upon disclosure, first, blood sugar status was not included since participants are required to offer their clinical data before checkup without the use of an invasive method such as "fingerstick" sampling to obtain the blood sugar level; second, the final pathological report of polyps was not illustrated because it was supposed that all polyps should be sampled for their nature to determine whether participants' potentially have colorectal polyps, which are considered to be precancerous lesions [3–6].

An association was observed between the colorectal polyp group and age, *Helicobacter pylori* infection, hypertension, and BMI regardless of gender ($p < 0.0001$). Colorectal polyps ($p = 0.0256$) and BMI (overweight, $p = 0.0111$) were significantly different among female patients. Age, *Helicobacter pylori* infection, and hypertension were common risk factors for colorectal polyps.

Regarding age, many studies have reported the association between age and colorectal polyps [29,30], suggesting that CRC screening should be performed around the age of 50–60 years in the general population owing to >80% of CRCs being diagnosed over the age of 60 years, which is consistent with our results [31–34].

Helicobacter pylori infection is highly associated with hyperplastic polyps [34–38], fundic gland polyps [34], and colorectal polyps [16,39–42]. Physiological mechanisms are still unclear, although Meira et al. [34] reported that *Helicobacter pylori* infection is associated with chronic inflammation-induced DNA damage and increased levels of serum gastrin, and *Helicobacter pylori* CagA status may be the cause of colonic neoplasm formation [43–46].

Metabolic syndrome is characterized by the presence of at least three of the following five factors—abdominal obesity, elevated triglyceride levels, decreased high-density lipoprotein cholesterol levels, hypertension, and high fasting glucose levels [47]—and contributes to various diseases, including gastric neoplasm and colorectal neoplasm [48]. In our study, hypertension and BMI were significant across genders in our analysis, and as mentioned before, noninvasive methods are available for easily obtaining factor data from individuals before endoscopy. In our study, hypertension and BMI were both significantly associated with the presence of colorectal polyps.

As discussed in [27], BMI statuses, both overweight and underweight, can alter gut metabolism, and as [28] pointed out, the change in metabolism significantly relates to colorectal cancer and polyps. We hypothesis that BMI is a significant indicator for different colorectal health; therefore, the risk factor might change from one BMI status to another. The results of AUC prove that our hypothesis is correct. For some BMI status, it is easier to identify the patient with colorectal polyps and others are not. The risk factors also differ from one BMI status to another.

The bulk of data has validated dental problems as a risk factor for colon neoplasm development [15,49]. We surmise that periodontal disease may induce chronic inflammation, resulting in immune dysregulation, and alters gut microbiota, which could be one possible pathway responsible for colorectal carcinogenesis [50–52]. It was also found that GB polyps/stones are also related to colorectal polyps, consistent with recent studies [17,53]. This may be attributed to GB polyps/stones and colorectal polyps that share some risk factors, such as obesity and metabolic syndrome [54].

In our study, there is no doubt that all aforementioned risk factors are noninvasive indicators of colorectal polyp formation [48]. Our risk stratification tool, which is built based on identified risk factors with a machine learning method, exhibits high sensitivities (70–80%) compared with that in noninvasive tools developed by previous studies (60–70%) [55]. Other decision tree-based studies [25,26] build noninvasive stratification tools using annual check-up data for non-alcoholic fatty liver obtained in AUC ranges from 85 to 87%. Compared with previous studies, the proposed model outperformed in several subgroups, such as elder obsessive individuals.

The limitations of this study were as follows: (1) its retrospective nature; (2) it was conducted at a single institution with a Taiwanese population; (3) our sigmoidoscopy is conducted under anaesthetization. Thus, our dataset excluded patients with BMI > 35 due to the protocol code of the anesthesiologist. Future researchers can build an RF model for this subgroup or collect data of non-anesthetized sigmoidoscopy diagnostics.

5. Conclusions

In this research, we proposed a new approach for building a risk stratification tool for colorectal polyps. First, we identified a set of promising risk factors using traditional statistical analysis such as z-test and logistics regression. We find that risk factors significantly differ for different genders, ages, and BMI statuses. Then, we separate patients with key demographic characteristics, which we believe each subgroup has a different set of risk factors. Then, we implement random forest to build a machine learning model to stratify patients with and without colorectal polyps. Colonoscopy verification is warranted in those

50 years of age or older, with hypertension, and infected with *Helicobacter pylori*. However, colonoscopy verification is warranted in individuals with tooth diseases and GB polyps.

For obese females, GB polyps warrant further colonoscopy verification. For males over age 50 and not underweight, hypertension is a strong indicator of possible colorectal polyps. We also find that for either underweight or obese patients, the AUC is higher than other groups. That is, abnormal weight is a strong indicator of health status, and different health statuses should be modeled differently. This is verified by our design of grouping patients with different demographic characteristics before building a machine learning model.

Our risk stratification tool can help healthcare centers identify patients who need further colonoscopy. This tool provides two major benefits: first, it helps clinicians conduct colonoscopy and discover precancerous lesions earlier to prevent cancer; second, it reduces the time and financial burden of healthcare centers in conducting unnecessary colonoscopies.

Author Contributions: C.L. conducted statistical analysis and created the machine learning algorithm and contributed to the writing of the manuscript and revised the manuscript according to reviewers' comments. C.-J.L. contributed to the implementation of the machine learning algorithm. T.-H.L. contributed to data collection, data cleaning, and manuscript writing. C.-F.K., B.C.-J.P. and H.-T.C. contributed to data cleaning, literature review, and identification of possible risk factors in this study. C.-C.L. helped with data collection. T.-H.C. provided initial ideas and research directions and finalized the manuscript. All authors contributed significantly to this study. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from Chang Gung Memorial Hospital, Taoyuan, Taiwan (CORPG3F0261).

Institutional Review Board Statement: This study was approved by the Ethics Committee of the Institutional Review Board of Chang Gung Memorial Hospital and conducted according to the ethical principles of the Declaration of Helsinki as reflected in the a priori approval by the institution's human research committee.

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to the protection of patients' privacy and restriction from the Ethics Committee of the Institutional Review Board of Chang Gung Memorial Hospital.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Torre, L.A.; Bray, F.; Siegel, R.L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. Global cancer statistics, 2012. *CA Cancer J. Clin.* **2015**, *65*, 87–108. [CrossRef] [PubMed]
2. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer Incidence and Mortality Worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, E359–E386. [CrossRef] [PubMed]
3. Calderwood, A.H.; Lasser, K.E.; Roy, H.K. Colon adenoma features and their impact on risk of future advanced adenomas and colorectal cancer. *World J. Gastrointest. Oncol.* **2016**, *8*, 826–834. [CrossRef] [PubMed]
4. Schmitz, J.M.; Stolte, M. Gastric Polyps as Precancerous Lesions. *Gastrointest. Endosc. Clin. N. Am.* **1997**, *7*, 29–46. [CrossRef]
5. Zheng, E.; Ni, S.; Yu, Y.; Wang, Y.; Weng, X.; Zheng, L. Impact of gender and age on the occurrence of gastric polyps: Data analysis of 69575 southeastern Chinese patients. *Turk. J. Gastroenterol.* **2015**, *26*, 474–479. [CrossRef]
6. Islam, R.S.; Patel, N.C.; Lam-Himlin, D.; Nguyen, C.C. Gastric Polyps: A Review of Clinical, Endoscopic, and Histopathologic Features and Management Decisions. *Gastroenterol. Hepatol.* **2013**, *9*, 640–651.
7. Citarda, F.; Tomaselli, G.; Capocaccia, R.; Barcherini, S.; Crespi, M. The Italian Multicentre Study Group Efficacy in standard clinical practice of colonoscopic polypectomy in reducing colorectal cancer incidence. *Gut* **2001**, *48*, 812–815. [CrossRef]
8. Carmack, S.W.; Genta, R.M.; Graham, D.Y.; Lauwers, G.Y. Management of gastric polyps: A pathology-based guide for gastroenterologists. *Nat. Rev. Gastroenterol. Hepatol.* **2009**, *6*, 331–341. [CrossRef]
9. Kim, S.Y.; Kim, H.-S.; Park, H.J. Adverse events related to colonoscopy: Global trends and future challenges. *World J. Gastroenterol.* **2019**, *25*, 190–204. [CrossRef]

10. Greenspan, M.; Prickett, E.; Melson, J. High Clinical Patient Workload Leads to Increased Premature Adenomatous Polyp Surveillance Colonoscopy. *Am. J. Gastroenterol.* **2015**, *110*, S601. [CrossRef]
11. Almadi, M.; Sewitch, M.; Barkun, A.N.; Martel, M.; Joseph, L. Adenoma Detection Rates Decline with Increasing Procedural Hours in an Endoscopist's Workload. *Can. J. Gastroenterol. Hepatol.* **2015**, *29*, 304–308. [CrossRef]
12. Sey, M.S.L.; Gregor, J.; Adams, P.; Khanna, N.; Vinden, C.; Driman, D.; Chande, N. Wait Times for Diagnostic Colonoscopy among Outpatients with Colorectal Cancer: A Comparison with Canadian Association of Gastroenterology Targets. *Can. J. Gastroenterol.* **2012**, *26*, 894–896. [CrossRef]
13. Cappell, M.S. The pathophysiology, clinical presentation, and diagnosis of colon cancer and adenomatous polyps. *Med Clin. N. Am.* **2005**, *89*, 1–42. [CrossRef]
14. Ren, H.G.; Luu, H.N.; Cai, H.; Xiang, Y.B.; Steinwandel, M.; Gao, Y.T.; Hargreaves, M.; Zheng, W.; Blot, W.J.; Long, J.R.; et al. Oral health and risk of colorectal cancer: Results from three cohort studies and a meta-analysis. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **2016**, *27*, 1329–1336. [CrossRef]
15. Momen-Heravi, F.; Babic, A.; Tworoger, S.S.; Zhang, L.; Wu, K.; Smith-Warner, S.A.; Ogino, S.; Chan, A.T.; Meyerhardt, J.; Giovannucci, E.; et al. Periodontal disease, tooth loss and colorectal cancer risk: Results from the Nurses' Health Study. *Int. J. Cancer* **2017**, *140*, 646–652. [CrossRef]
16. Brim, H.; Zahaf, M.; Laiyemo, A.O.; Nouraie, M.; Pérez-Pérez, G.I.; Smoot, D.T.; Lee, E.; Razjouyan, H.; Ashktorab, H. Gastric Helicobacter pylori infection associates with an increased risk of colorectal polyps in African Americans. *BMC Cancer* **2014**, *14*, 296. [CrossRef]
17. Liu, Y.L.; Wu, J.S.; Yang, Y.C.; Lu, F.H.; Lee, C.T.; Lin, W.J.; Chang, C.J. Gallbladder stones and gallbladder polyps associated with increased risk of colorectal adenoma in men. *J. Gastroenterol. Hepatol.* **2018**, *33*, 800–806. [CrossRef]
18. Xiao, S.; Zhou, L. Gastric cancer: Metabolic and metabolomics perspectives (Review). *Int. J. Oncol.* **2017**, *51*, 5–17. [CrossRef]
19. Ford, I.; Robertson, M.; Komajda, M.; Böhm, M.; Borer, J.S.; Tavazzi, L.; Swedberg, K. Top ten risk factors for morbidity and mortality in patients with chronic systolic heart failure and elevated heart rate: The SHIFT Risk Model. *Int. J. Cardiol.* **2015**, *184*, 163–169. [CrossRef]
20. Okada, H.; Fukui, M.; Tanaka, M.; Matsumoto, S.; Mineoka, Y.; Nakanishi, N.; Asano, M.; Yamazaki, M.; Hasegawa, G.; Nakamura, N. Visit-to-Visit Blood Pressure Variability Is a Novel Risk Factor for the Development and Progression of Diabetic Nephropathy in Patients with Type 2 Diabetes. *Diabetes Care* **2013**, *36*, 1908–1912. [CrossRef]
21. Khalilia, M.; Chakraborty, S.; Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform. Decis. Mak.* **2011**, *11*, 51. [CrossRef]
22. Graham, D.Y.; Miftahussurur, M. Helicobacter pylori urease for diagnosis of Helicobacter pylori infection: A mini review. *J. Adv. Res.* **2018**, *13*, 51–57. [CrossRef]
23. Andren-Sandberg, A. Diagnosis and management of gallbladder polyps. *N. Am. J. Med. Sci.* **2012**, *4*, 203–211. [CrossRef]
24. Abdulqader, D.M.; Abdulazeez, A.M.; Zeebaree, D.Q. Machine learning supervised algorithms of gene selection: A review. *Mach. Learn.* **2020**, *62*, 233–244.
25. Goldman, O.; Ben-Assuli, O.; Rogowski, O.; Zeltser, D.; Shapira, I.; Berliner, S.; Zelber-Sagi, S.; Shenhar-Tsarfaty, S. Non-alcoholic Fatty Liver and Liver Fibrosis Predictive Analytics: Risk Prediction and Machine Learning Techniques for Improved Preventive Medicine. *J. Med. Syst.* **2021**, *45*, 22. [CrossRef]
26. Fialoke, S.; Malarstig, A.; Miller, M.R.; Dumitriu, A. Application of Machine Learning Methods to Predict Non-Alcoholic Steatohepatitis (NASH) in Non-Alcoholic Fatty Liver (NAFL) Patients. *AMIA Annu. Symp. Proc.* **2018**, *2018*, 430–439.
27. Wan, Y.; Yuan, J.; Li, J.; Li, H.; Yin, K.; Wang, F.; Li, D. Overweight and underweight status are linked to specific gut microbiota and intestinal tricarboxylic acid cycle intermediates. *Clin. Nutr.* **2020**, *39*, 3189–3198. [CrossRef]
28. Jain, R.; Pickens, C.A.; Fenton, J.I. The role of the lipidome in obesity-mediated colon cancer risk. *J. Nutr. Biochem.* **2018**, *59*, 1–9. [CrossRef]
29. Cao, W.; Hou, G.; Zhang, X.; San, H.; Zheng, J. Potential risk factors related to the development of gastric polyps. *Immunopharmacol. Immunotoxicol.* **2018**, *40*, 338–343. [CrossRef]
30. Chen, H.; Li, N.; Ren, J.; Feng, X.; Lyu, Z.; Wei, L.; Li, X.; Guo, L.; Zheng, Z.; Zou, S.; et al. Participation and yield of a population-based colorectal cancer screening programme in China. *Gut* **2018**, *68*, 1450–1457. [CrossRef] [PubMed]
31. Hussein Kamareddine, M.; Ghosn, Y.; Karam, K.; Nader, A.A.; El-Mahmoud, A.; Bou-Ayash, N.; El-Khoury, M.; Farhat, S. Adenoma Detection before and after the age of 50: A retrospective analysis of Lebanese outpatients. *BMJ Open Gastroenterol.* **2018**, *5*, 000253. [CrossRef] [PubMed]
32. Wolf, A.M.D.; Fontham, E.T.H.; Church, T.R.; Flowers, C.R.; Guerra, C.E.; LaMonte, S.J.; Etzioni, R.; McKenna, M.T.; Oeffinger, K.C.; Shih, Y.-C.T.; et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J. Clin.* **2018**, *68*, 250–281. [CrossRef] [PubMed]
33. Schreuders, E.H.; Ruco, A.; Rabeneck, L.; Schoen, R.E.; Sung, J.J.Y.; Young, G.; Kuipers, E.J. Colorectal cancer screening: A global overview of existing programmes. *Gut* **2015**, *64*, 1637–1649. [CrossRef] [PubMed]
34. Bevan, R.; Rutter, M.D. Colorectal Cancer Screening-Who, How, and When? *Clin. Endosc.* **2018**, *51*, 37–49. [CrossRef]
35. Kang, K.H.; Hwang, S.H.; Kim, N.; Kim, D.-H.; Kim, S.Y.; Hyun, J.J.; Jung, S.W.; Koo, J.S.; Jung, Y.K.; Yim, H.J.; et al. The Effect of Helicobacter pylori Infection on Recurrence of Gastric Hyperplastic Polyp after Endoscopic Removal. *Korean J. Gastroenterol.* **2018**, *71*, 213–218. [CrossRef]

36. Anjiki, H.; Mukaisho, K.-I.; Kadomoto, Y.; Doi, H.; Yoshikawa, K.; Nakayama, T.; Vo, D.T.-N.; Hattori, T.; Sugihara, H. Adenocarcinoma arising in multiple hyperplastic polyps in a patient with Helicobacter pylori infection and hypergastrinemia during long-term proton pump inhibitor therapy. *Clin. J. Gastroenterol.* **2017**, *10*, 128–136. [CrossRef]
37. Markowski, A.R.; Markowska, A.; Guzinska-Ustymowicz, K. Pathophysiological and clinical aspects of gastric hyperplastic polyps. *World J. Gastroenterol.* **2016**, *22*, 8883–8891. [CrossRef]
38. Togo, K.; Ueo, T.; Yonemasu, H.; Honda, H.; Ishida, T.; Tanabe, H.; Yao, K.; Iwashita, A.; Murakami, K. Two cases of adenocarcinoma occurring in sporadic fundic gland polyps observed by magnifying endoscopy with narrow band imaging. *World J. Gastroenterol.* **2016**, *22*, 9028–9034. [CrossRef]
39. Tongtawee, T.; Simawaranon, T.; Wattanawongdon, W. Role of screening colonoscopy for colorectal tumors in Helicobacter pylori-related chronic gastritis with MDM2 SNP309 G/G homozygous: A prospective cross-sectional study in Thailand. *Turk. J. Gastroenterol.* **2018**, *29*, 555–560. [CrossRef]
40. Kumar, A.; Kim, M.; Lukin, D.J. Helicobacter pylori is associated with increased risk of serrated colonic polyps: Analysis of serrated polyp risk factors. *Indian J. Gastroenterol.* **2018**, *37*, 235–242. [CrossRef]
41. Nam, J.H.; Hong, C.W.; Kim, B.C.; Shin, A.; Ryu, K.H.; Park, B.J.; Kim, B.; Sohn, D.K.; Han, K.S.; Kim, J.; et al. Helicobacter pylori infection is an independent risk factor for colonic adenomatous neoplasms. *Cancer Causes Control.* **2017**, *28*, 107–115. [CrossRef]
42. Meira, L.B.; Bugni, J.M.; Green, S.L.; Lee, C.-W.; Pang, B.; Borenshtein, D.; Rickman, B.H.; Rogers, A.B.; Moroski-Erkul, C.A.; McFaline, J.L.; et al. DNA damage induced by chronic inflammation contributes to colon carcinogenesis in mice. *J. Clin. Investig.* **2008**, *118*, 2516–2525. [CrossRef]
43. Thorburn, C.M.; Friedman, G.D.; Dickinson, C.J.; Vogelman, J.H.; Orentreich, N.; Parsonnet, J. Gastrin and colorectal cancer: A prospective study. *Gastroenterology* **1998**, *115*, 275–280. [CrossRef]
44. Georgopoulos, S.D.; Polymeros, D.; Triantafyllou, K.; Spiliadi, C.; Mentis, A.; Karamanolis, D.G.; Ladas, S.D. Hypergastrinemia Is Associated with Increased Risk of Distal Colon Adenomas. *Digestion* **2006**, *74*, 42–46. [CrossRef]
45. Epplein, M.; Pawlita, M.; Michel, A.; Peek, R.M.; Cai, Q.; Blot, W.J. Helicobacter pylori Protein–Specific Antibodies and Risk of Colorectal Cancer. *Cancer Epidemiol. Biomark. Prev.* **2013**, *22*, 1964–1974. [CrossRef]
46. Shmuely, H.; Passaro, D.; Figer, A.; Niv, Y.; Pitlik, S.; Samra, Z.; Koren, R.; Yahav, J. Relationship between Helicobacter pylori CagA status and colorectal cancer. *Am. J. Gastroenterol.* **2001**, *96*, 3406–3410. [CrossRef]
47. Grundy, S.M.; Brewer, H.B.; Cleeman, J.I., Jr.; Smith, S.C., Jr.; Lenfant, C. Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation* **2004**, *109*, 433–438. [CrossRef]
48. Park, W.; Lee, H.; Kim, E.H.; Yoon, J.Y.; Park, J.C.; Shin, S.K.; Kil Lee, S.; Lee, Y.C.; Kim, W.H.; Noh, S.H. Metabolic syndrome is an independent risk factor for synchronous colorectal neoplasm in patients with gastric neoplasm. *J. Gastroenterol. Hepatol.* **2012**, *27*, 1490–1497. [CrossRef]
49. Chou, S.H.; Tung, Y.C.; Wu, L.S.; Chang, C.J.; Kung, S.; Chu, P.H. Severity of chronic periodontitis and risk of gastrointestinal cancers: A population-based follow-up study from Taiwan. *Medicine* **2018**, *97*, e11386. [CrossRef]
50. Lauritano, D.; Sbordone, L.; Nardone, M.; Iapichino, A.; Scapoli, L.; Carinci, F. Focus on periodontal disease and colorectal carcinoma. *Oral Implant.* **2017**, *10*, 229–233. [CrossRef]
51. Gao, Z.; Guo, B.; Gao, R.; Zhu, Q.; Qin, H. Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol.* **2015**, *6*, 20. [CrossRef] [PubMed]
52. Moutsopoulos, N.M.; Madianos, P.N. Low-Grade Inflammation in Chronic Infectious Diseases: Paradigm of Periodontal Infections. *Ann. N. Y. Acad. Sci.* **2006**, *1088*, 251–264. [CrossRef] [PubMed]
53. Stergios, K.; Damaskos, C.; Frountzas, M.; Nikiteas, N.; Lalude, O. Can gallbladder polyps predict colorectal adenoma or even neoplasia? A systematic review. *Int. J. Surg.* **2016**, *33*, 23–27. [CrossRef] [PubMed]
54. Lim, S.H.; Kim, D.H.; Park, M.J.; Kim, Y.S.; Kim, C.H.; Yim, J.Y.; Cho, K.R.; Kim, S.S.; Choi, S.H.; Kim, N.; et al. Is Metabolic Syndrome One of the Risk Factors for Gallbladder Polyps Found by Ultrasonography during Health Screening? *Gut Liver* **2007**, *1*, 138–144. [CrossRef]
55. Tanwar, S.; Vijayalakshmi, S. Comparative Analysis and Proposal of Deep Learning Based Colorectal Cancer Polyps Classification Technique. *J. Comput. Theor. Nanosci.* **2020**, *17*, 2354–2362. [CrossRef]