MASHINALI O‘QITISH ASOSIDA VEB SERVERLARDA ZAIFLIKLARNI ANIQLASH MODELI

Barotova Zahro Akmaljon qizi
Toshkent, O‘zbekiston
Zbarotova04@gmail.com

Annotatsiya: Tarmoq zaifliklaridan kelib chiqadigan xavfsizlik muammolarini hal qilish uchun veb ilovalarning saytlararo skript hujumlarini samarali oldini olish va tarmoq xavfsizligi bilan bog’liq hodisalarni kamaytirish uchun mashinani o‘rganishga asoslangan veb ilovalarning zaifligini aniqlash usuli taklif etiladi. Mavjud xavfsizlik zaifligini aniqlash texnologiyasini chuqur o‘rganish orqali, mashinani o‘rganish xavfsizligi zaifligini aniqlash texnologiyasini ishlab chiqish jarayoni bilan birgalikda, xavfsizlik zaifligini aniqlash modelining talablari batafsil tahlil qilinadi va veb-saytlar uchun xavfsizlik zaifligini aniqlash modelini batafsil tahlil qiladi. dastur ishlab chiqilgan va amalga oshiriladi. Mavjud tarmoq zaifligini aniqlash texnologiyasi va vositalariga asoslanib, tekshirish kodini identifikatsiyalash funksiyasi qo‘shildi, bu ma’lumotlarni serverga faqat tasdiqlash kodini kiritish orqali yuborish mumkin bo‘lgan muammoni hal qiladi. Serverni filtrlash qoidalariga ko‘ra, server filtrlashni chetlab o‘tuvchi tarmoq kodi tuzilgan. Eksperimental natijalar shuni ko‘rsatadiki, modelda o‘tkazib yuborilgan aniqlash va noto‘g’ri signalizatsiya darajasi past va takomillashtirilgan model yanada samaraliroq.

Kalit so‘zlar: mashinali o‘qitish, veb zaifliklar, parametr ko‘rsatkichlari.

I. Kirish.

Veb-ilovalar bank, moliya, marketing, soliq, ijtimoiy o‘zaro aloqalar va hatto tibbiyot kabi sohalarda keng qo‘llanilishi tufayli kundalik hayotimizning asosiy qismiga aylandi. Ko‘pgina muhim operatsiyalar Internetda amalga oshiriladi, ular shaxsiy ma’lumotlar bo‘lishi mumkin bo‘lgan identifikatsiya ma’lumotlari, parollar, pin yoki parollar, tibbiy ma’lumotlar yoki tranzaksiyalar, xarid tafsilotlari, soliq

deklaratsiyasi va boshqalar kabi moliyaviy tafsilotlar bo‘lishi mumkin. Garchi veb-ilovalar xavfsizlik va xavfsizlik choralarini ta’minlaydi. himoya qilish uchun ba’zi veb zaifliklar mavjud bo‘lib, ular mijoz tomonidan yoki server tomonida raqib tomonidan foydalaniladigan bo‘shliqlar, xatolar yoki bo‘shliqlardir. Darhaqiqat, xabar qilingan veb-zaifliklar soni [6], va vebga asoslangan hujumlar chastotasi kundan-kunga sezilarli darajada oshib bormoqda, bu esa kuchliroq xavfsizlikni taqozo etadi. Biroq, veb-ilovalar juda murakkab va ularni ishlab chiqishda qo‘llaniladigan turli xil murakkab dasturlash texnikasi tufayli tahlil qilish qiyin. Shu sababli, ushbu veb zaifliklarini aniqlash to‘g’ridan-to‘g’ri vazifa emas va veb-ilovani himoya qilishning muhim tarkibiy qismlaridan biridir.

II. Mashinali o‘qitishga asoslangan model

Yangi etiketli ma’lumotlar to‘plamini tezda yaratish uchun bir nechta turli usullardan foydalanildi. Dastlab OWASP va MITER resurslaridan zaiflik namunalarini to‘plandi. Bularga ko‘rgazmali misollar, shuningdek hujjatlashtirilgan real ilovalarning zaifliklarini misol qilsa bo‘ladi. Keyin so‘rov va javob juftliklari minimallashtirilgan matn formatiga aylantiriladi va shablonimizga mos keladigan kontekstual ma’lumotlarga o‘tkaziladi. Shuningdek, zaif misollar asosida misollarni himoyasiz qilish uchun kichik o‘zgarishlar bilan salbiy namunalar yaratildi. Keyinchalik, vazifa tavsifini, javob uchun shablonni va ICL uchun yorliqli misollar ro‘yxatini o‘z ichiga olgan bir nechta takliflar yaratiladi. Ushbu matndan foydalanib, OpenAI tomonidan gpt3.5-turbo modelini taklif qildik. Asosiy model GPT3-ga asoslangan LLM bo‘lib, suhbatlashish uchun sozlangan va kechikish uchun optimallashtirilgan. Ko‘p yuzlab milliardlab matn tokenlari bilan o‘qitilgan, u CoQA mezonida bir necha marta o‘tkazilganda 85 F1 ball oldi. Ma’lumotni yaratish usulimizda biz ICL ga hissa qo‘shadigan ilg’or tezkor muhandislik texnikasi bilan birgalikda bir necha marta chaqiruvdan foydalanamiz. Kirish va chiqish uchun ruxsat etilgan maksimal ketma-ketlik uzunligi 4096 ta token. Bu bizga kengaytirilgan kontekstga ega bo‘lgan uzunroq so‘rovlardan foydalanish imkonini beradi. So‘rovning sifati generativ modellarning ishlashiga bevosita ta’sir qiladi [1].

Namuna olish. 1-jadvalda ko‘rinib turganidek, ma’lumotlar to‘plami qatlamli tanlama yordamida bir nechta kichikroq kichik ma’lumotlar to‘plamlariga tanlanadi, “XLarge” ma’lumotlar to‘plami bundan mustasno. Ma’lumotlar to‘plamlari o‘zaro tekshirish uchun mo‘ljallangan, agar 10 burmaga bo‘lingan bo‘lsa, har bir qavat teng miqdordagi sinflarga ega bo‘ladi. “XLarge” ma’lumotlar to‘plami asosan salbiy namunalarga ega bo‘lgan barcha etiketli ma’lumotlar to‘plamidan foydalangan holda yaratilgan. Ushbu ma’lumotlar to‘plami, shuningdek, 10 marta bo‘linish uchun mo‘ljallangan, ularning har birida teng miqdordagi teglar mavjud.

1-jadval

Ma’lumotlar to‘plami va ularning xususiyatlari

Nomi	Jami	Zaiflik emas	CWE-639	CWE-209
Small	150	50	50	50
Base	300	100	100	100
Large	600	200	200	200
XLarge	1780	1340	200	200

Eksperimental loyihalash. Eksperimental loyihalash ikki bosqichga bo‘linadi: ma’lumotlarni tayyorlash va modelni tekshirish. Birinchi bosqichning maqsadi ma’lumotlarga oid barcha bosqichlarni, jumladan, ma’lumotlarni yig’ish, tozalash, yaratish va tayyorlashni o‘z ichiga olgan o‘quv ma’lumotlar to‘plamini yaratishdir. Ikkinchi bosqich oldingi bosqichda yaratilgan ma’lumotlar to‘plamidan foydalanadi va baholash uchun ishlatiladigan tasniflash uchun transformator modelini o‘rgatadi.

1-qadam. Zaiflik namunalarini to‘plash

2-qadam. Namunalardan bir necha marta ko‘rsatmalar yaratish

3-qadam. Ma’lumotlar to‘plamini yaratish uchun GPT3 bilan ko‘rsatmalardan foydalanish

4-qadam. O‘quv ma’lumotlar to‘plamini tozalash, belgilash va tayyorlash

5-qadam. SetFit yordamida LLMlarni nozik sozlashni testlash va sinab ko‘rish

6-qadam Tahlil qilish

Giperparametrlar. Giperparametrlar modelni o‘qitish jarayonini sozlaydigan yuqori darajadagi o‘zgaruvchilardir. Ular modelning ishlashi va harakatini nazorat qiladi. Tadqiqot uchun eng yaxshi parametrlarni tanlash juda muhim, chunki suboptimal parametrlar yakuniy natijalarni chalg’itishi mumkin. Giperparametrlarni hisoblash mumkin emas va optimal konfiguratsiyani topish uchun tajriba talab etiladi. Parametr konfiguratsiyasining katta hajmini takrorlash va ularni eng yaxshi ishlash uchun sinovdan o‘tkazish adabiyotda “giperparametrlarni sozlash” yoki “giperparametrlarni optimallashtirish” deb nomlanishi mumkin. Tadqiqot ishida modelning konvergensiya tezligi va ishlashi uchun optimallashtiriladi. Cheklangan resurslar tufayli optimal giperparametrlarni qidirish maydonini cheklashimiz kerak edi. Bundan tashqari, umumiy eksperimentni 20 tagacha hisoblashni o‘rnatdik, bu tadqiqotning boshqa qismlari uchun vaqtni tejaydi. Tadqiqotda foydalanilgan yakuniy parametrlarni 2-jadvalda ko‘rish mumkin.

2-jadval

Trening uchun ishlatiladigan giperparametr konfiguratsiyasi

Parametr	**Qiymat**	**Ta’rifi**
Learning rate	2e-6	Model o‘z parametrlarini moslashtiradigan qadam o‘lchami
Epoch	3	Barcha ma’lumotlar to‘plami model orqali o‘tish soni
Iteratsiya sanog’i	20	Yaratiladigan matn juftlari soni
Seed (urug’)	25	Natijalarning takrorlanishini ta’minlash uchun tasodifiy urug’lik qiymati
Warmup nisbati	0.1	O‘qitish tezligini bosqichma-bosqich oshirishga bag’ishlangan o‘quv bosqichlarining nisbati
Partiya hajmi	32	Bitta iteratsiyada qayta ishlangan trening misollari soni

Baholash ko‘rsatkichlari. Taklif etilgan usulning samaradorligini baholash uchun biz o‘zaro tekshirishdan foydalanamiz. Mashinali o‘qitishda bashorat qilish modellarining samaradorligini baholashning mashhur usuli K-katta o‘zaro tekshirish hisoblanadi. Ma’lumotlar tasodifiy sonlar generatori yordamida bir xil o‘lchamdagi K burmalarga bo‘linadi. Keyin qolgan K-1 burmalari modelni o‘rgatish uchun ishlatiladi, K burmalardan biri har bir iteratsiyada tasdiqlash sifatida xizmat qiladi. O‘quv majmuasida model ishlab chiqiladi va tekshirish to‘plami uning ishlashini baholash uchun ishlatiladi. K burmalarning har biri tekshirish to‘plami sifatida aniq bir marta ishlatiladi va protsedura K marta takrorlanadi. Oxir-oqibat, har bir iteratsiya natijalarini o‘rtacha hisoblash orqali modelning ishlashining umumiy bahosi yaratiladi. Ma’lumotlar to‘plamining o‘lchami va ishlashini baholash uchun zarur bo‘lgan aniqlik darajasi K ni tanlashga ta’sir qiladi, bu bizning tadqiqotimizda 10 ga teng.

An’anaviy ushlab turish yondashuvlari bilan solishtirganda, k-fold o‘zaro tekshirishning asosiy afzalligi modelning ishlashini yanada chuqurroq tekshirish imkonini beradi. U ma’lumotlarning har qanday o‘ziga xosligi yoki tasodifiy tebranishlarining ta’sirini kamaytiradigan ma’lumotlarning turli kichik to‘plamlarida o‘rgatish va sinovdan o‘tkazish orqali modelning samaradorligini yanada ishonchli baholashi mumkin. Har bir iteratsiyada tekshirish to‘plamida modelning ishlashini baholash orqali k-fold o‘zaro tekshirish neyron tarmoqdagi yashirin qatlamlar soni kabi model giperparametrlarini nozik sozlash uchun ham ishlatilishi mumkin.

Ushbu muammoni hal qilish uchun biz tabaqalashtirilgan namunalardan foydalanamiz. Bu maqsadli o‘zgaruvchining taqsimlanishi namunaviy ma’lumotlarda yaxshi ifodalanishini ta’minlash uchun statistika va mashinali o‘qitishda qo‘llaniladigan namuna olish usulidir. Texnika maqsadli o‘zgaruvchiga qarab populatsyani kichik guruhlarga yoki qatlamlarga bo‘lish va keyin har bir qatlamdan uning hajmiga mutanosib ravishda tasodifiy tanlashni o‘z ichiga oladi. Bu har bir qatlamning namunada ifodalanishini va namuna taqsimoti populyatsiya taqsimotiga o‘xshashligini ta’minlaydi.

O‘qituvchili o‘qitish algoritmi bilan statistik tasnifni chalkashlik matritsasi bilan baholash mumkin. Tasdiqlash to‘plami bo‘yicha bashorat qilinganidan keyin chalkashlik matritsasi yaratiladi va to‘rtta natijani tavsiflaydi:

- True positive (TP) kuzatuv hujum bo‘lishi to‘g’ri bashorat qilingan.
- True negative (TN) kuzatuvi hujum bo‘lmasligi to‘g’ri bashorat qilingan.
- False positive (FP) kuzatuvi noto‘g’ri ravishda hujum deb taxmin qilingan.
- False negative (FN) kuzatuvi hujum bo‘lmasligi noto‘g’ri prognoz qilingan.

Tasniflash hisoboti - bu asosiy tasnif ko‘rsatkichlarini ko‘rsatadigan matnli hisobot hisoblanib, Aniqlik (Accuracy), aniqlik (Precision), eslab qolish (Recall) va F1 Score va MCCga bo‘linadi.

Biz foydalanadigan baholash ko‘rsatkichlari ko‘pincha ikkilik tasniflash masalalarida qo‘llaniladi va mashinali o‘qitish modellarini to‘liq baholashni taklif qiladi. Biz barcha yorliq sinflarini birgalikda baholash uchun makro o‘rtacha strategiyasidan foydalanamiz. Noto‘g’ri ijobiy prognozlar sezilarli xarajatlarga ega bo‘lgan holatlarda, aniqlik barcha ijobiy prognozlar orasida aniq ijobiy prognozlarning ulushini baholaydi. To‘g’ri aniqlangan haqiqiy ijobiylarning ulushi eslab qolish bilan o‘lchanadi, bu noto‘g’ri salbiy bashoratlar qimmatga tushganda hal qiluvchi ko‘rsatkichdir. Ushbu ko‘rsatkichlarning ikkalasi ham oraliq qiymatlar bo‘lib, modelning umumiy ishlashini yaxshi ko‘rsatmaydi. Aniqlik va eslab qolishni idrok etish va tushunish oson bo‘lsada, agar ma’lumotlar to‘plami muvozanatsiz bo‘lsa yoki noto‘g’ri musbat va noto‘g’ri salbiylarning narxi teng bo‘lmasa, ular aldamchi bo‘lishi mumkin. F1 balli bu holatda foydalanish uchun ajoyib statistik hisoblanadi, chunki u aniqlik va eslab qolishning garmonik o‘rtachasidir[2].

(1)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Aniqlik hujumlar (haqiqiy ijobiy) sifatida to‘g’ri tasniflangan kuzatuvlarning ulushi sifatida aniqlanadi.

(2)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall hujum sifatida tasniflangan hujumlarning ulushi sifatida aniqlanadi.

(3)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 score aniqlik va eslab qolishning garmonik o‘rtacha ko‘rsatkichidir.

(4)

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

(5)

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TN} + \text{FN})}}$$

Modelning umumiy ishlashini aks ettiruvchi ko‘rsatkichga ega bo‘lish juda muhim, shu jumladan salbiy statistika, ayniqsa noto‘g’ri optimistik prognozlar jiddiy oqibatlarga olib kelishi mumkin bo‘lgan holatlarda. Shunday qilib, biz optimallashtirish uchun Metyu korrelyatsiya koeffitsientini (MCC) tanladik. MCC ham to‘g’ri, ham noto‘g’ri ijobiy va salbiylarni hisobga oladi va bashorat qilingan va haqiqiy teglar o‘rtasidagi korrelyatsiyani hisoblaydi. Balanssiz ma’lumotlar to‘plamida aldamchi bo‘lishi mumkin bo‘lgan aniqlik kabi boshqa baholash ko‘rsatkichlaridan farqli o‘laroq, MCC ma’lumotlar to‘plamida ijobiy va salbiy holatlarning taqsimlanishini ko‘rib chiqadi. MCCdan foydalanishning yana bir afzalligi shundaki, u bir nechta modellarning ishlashini taqqoslashni osonlashtiradigan modelning ishlashini umumlashtiruvchi bitta skaler raqamni ishlab chiqaradi. Bundan tashqari, qaror qabul qilish chegarasidagi o‘zgarishlar MCCga ta’sir qilmaydi, bu aniqlik va eslab qolish kabi baholash ko‘rsatkichlari uchun tashvish tug’dirishi mumkin. Natijada, tadqiqot MCCdan baholash statistikasi sifatida foydalanish orqali yanada to‘liqroq va qat’iy ko‘rib chiqilishi mumkin.

Kodning murakkabligi ko‘rsatkichlari. Kod sifatini ta’minlash zaifliklarning oldini olishning asosidir. Cheklangan miqdordagi zaifliklar noto‘g’ri ramka dizayni tufayli yuzaga kelganiga qaramay, ularning aksariyati kod darajasiga bog’liq. Ushbu tadqiqotda statik tahlilni qo‘llash orqali ob’yektiv kod sifatini ifodalash uchun murakkablik ko‘rsatkichlari yig’iladi. Bir nechta ko‘rsatkichlar to‘plangan va ushbu ko‘rsatkichlarning ta’riflari quyida ko‘rsatilgan:

- path: indeks sifatida foydalaniladigan fayl yo‘li;
- sloc: kodlarning jismoniy qatori, bu modul yoki funksiyadagi qatorlar soni;
- cyclomatic: dastur oqimini boshqarish grafigidagi davrlar soni;
- siklomatik zichlik: kodning mantiqiy qatorlaridagi sikllarning ulushi;
- operator: operatorlarning umumiy soni;
- operand: aniq operandlar soni;
- vocabulary: alohida operatorlar va operandlar yig’indisi;
- length: operatorlar va operandlarning paydo bo‘lish vaqtlari yig’indisi;
- difficulty: $D = \frac{\eta 1}{2} * \frac{N2}{\eta 2}$.
- maintainability (barqarorlik): kodning mantiqiy qatorlari, siklomatik murakkablik va Halstead harakatlaridan kelib chiqqan.

Qabul qiluvchining operatsion xarakteristikasi egri chizig’i (ROC) turli chegaralarda tasniflash uchun ishlash o‘lchovidir. ROC bu ehtimollik egri chizig’i bo‘lib, bu yerda AUC ajralish darajasini ifodalaydi, bu model sinflarni qanchalik yaxshi farqlay olishini tavsiflaydi. 2.6-rasmda ko‘rsatilganidek, ROC odatda Y o‘qi bo‘yicha haqiqiy musbat tezlikni va X o‘qida noto‘g’ri musbat tezlikni ko‘rsatadi, bu “ideal” nuqta uchastkaning yuqori chap burchagi ekanligini anglatadi - nol va noto‘g’ri musbat ko‘rsatkichi va birning haqiqiy ijobiy darajasi. ROCning “tikligi” ham muhimdir, chunki noto‘g’ri musbat ko‘rsatkichni minimallashtirgan holda haqiqiy ijobiy ko‘rsatkichni maksimal darajada oshirish idealdir.

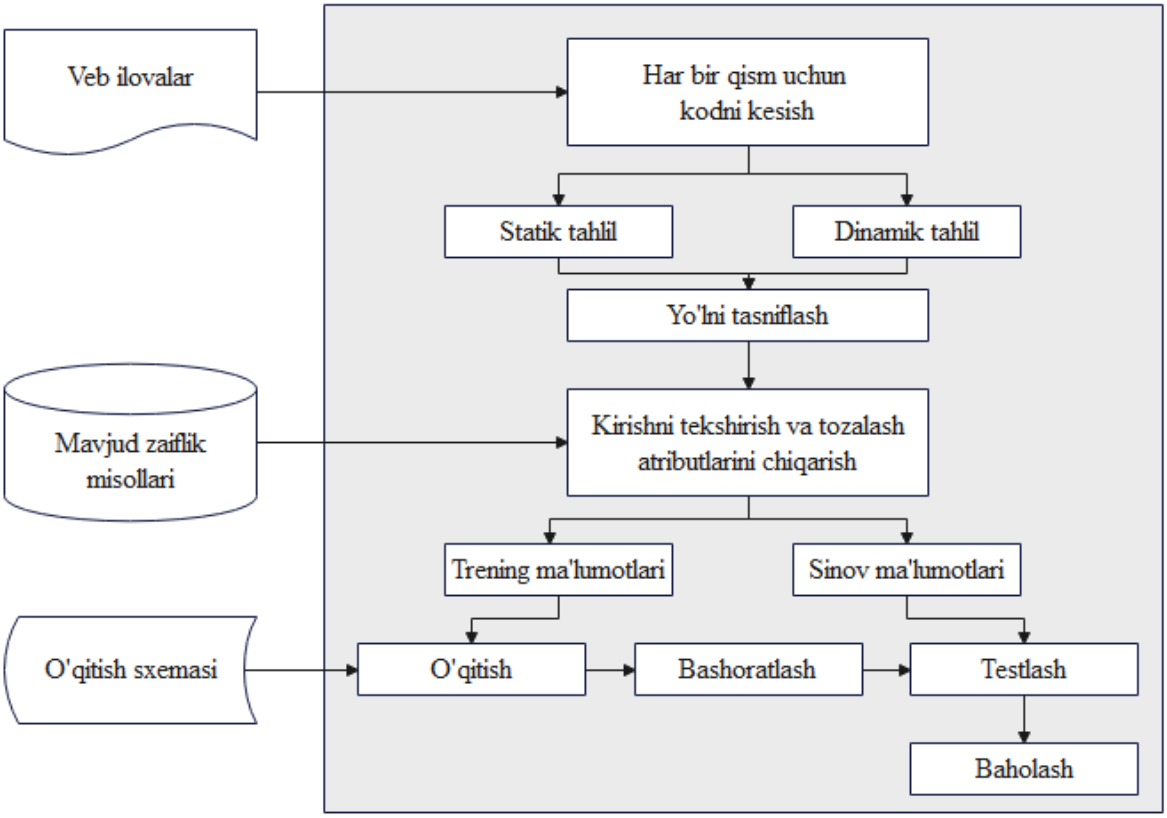
Taint asosidagi tahlilga asoslangan aniqlash modeli. Butun kontekst fayli haqida umumiy ma’lumotga ega bo‘lishlariga qaramay, tadqiqotchilar mavjud zaiflikning joylashuvini aniq aniqlash uchun ichidagi har bir bayonotni o‘rganib, manba kodini

chuqurroq ko‘rib chiqdilar. Shar birinchi bo‘lib o‘z ishini [3] XSS zaifligini aniqlash va avtomatik ravishda olib tashlash usulini taklif qilishdan boshladi. Ularning ishi asosan ikki xil bosqichga bo‘linishi mumkin. Potensial XSS zaifligini aniqlash uchun ular zaifliklarni aniqlash bo‘yicha odatiy statik tahlil bo‘lgan noravshanlanishga asoslangan tahlil usulidan foydalanganlar. Ikkinchi bosqichda ular XSS zaifligini olib tashlash uchun ikkita qadamni ishlab chiqdilar. Birinchidan, HTML chiqish bayonotida ishonchsiz ma’lumotlarga havola qilingan bayonotlarni aniqlash, mo‘ljallangan HTML chiqishlari va xavfsizlik jihatlariga ta’sir qilmasdan qochib qutulishi mumkin. Keyin manba kodidan har bir ishonchsiz ma’lumotni o‘rab turgan HTML hujjat tuzilishini ajratib olish va HTML kontekstini aniqlash uchun naqsh moslashuvidan foydalanish. Ikkinchidan, OWASP [4] tomonidan yaratilgan ESAPI-ning qochib ketgan APIlaridan foydalanib, asl kodni almashtirish sifatida xavfsiz kod tuzilmalarini yaratadi va bu jarayon to‘liq avtomatiklashtiriladi.

Biroq, dastlabki kodning zaif yoki zaif emasligini aniqlash uchun oldindan belgilangan qoidalarga asoslangan noravshanlikka asoslangan statik tahlil amaliy foydalanish paytida katta miqdordagi noto‘g’ri ijobiy natijalarga olib kelishi mumkin.

Gibrid tahlilga asoslangan aniqlash modeli. Filippning maqolasi [5] zararli qismni aniqlash uchun noravshanlikka asoslangan dinamik usulni joriy qildi. Ma’lumotlarning noravshanlanishi ishonchsiz ma’lumotlarni zararli ma’lumotlar sifatida belgilashdan boshlanadi, so‘ngra dastur orqali tarqaladi va tarqalishini kuzatib boradi. Ushbu protseduraga rioya qilish orqali, an’anaviy server tomoni himoya mexanizmlari buzilgan zararli ma’lumotlardan foydalanishni muvaffaqiyatli oldini olishi mumkin. Server tomonidan qo‘llaniladigan an’anaviy noravshanlikka asoslangan yondashuvlardan farqli o‘laroq, Filipp Firefox veb brauzerini o‘zgartirish orqali mijoz tomonidan yondashuvni taklif qildi. Shu bilan birga, u dinamik usullarni barcha turdagi nazorat bog’liqliklarini aniqlash uchun ishlatib bo‘lmasligini ta’kidladi, shuning uchun ularning “dinamik” ifloslanish usuli aslida XSS hujumlaridan to‘liq himoyaga erishish uchun statik va dinamik usullarning aralashmasidir. Shunday qilib, Balzarottining [6] qog’oziga asoslanib, Shar manba kodidan xususiyatlarni to‘plash uchun gibrid tahlilni

joriy etish orqali in’yeksiya hujumining oldini olish uchun yangi bashorat modelini ilgari surdi. Xususiyatlarni tanlash va tasniflashda ular o‘rganayotgan bog’liqlik grafigidagi har bir tugun "cho‘kish" edi. Sink - bu ma’lumotlar bazasi yoki veb mijoz bilan o‘zaro ta’sir qiladigan dastur bayonoti. Statik va dinamik tahliliga ko‘ra, ular ma’lumotlarga bog’liqlik grafigidan 22 ta atributni ajratib olishadi va ulardan klassifikatorlarni o‘qitish va sinab ko‘rish uchun kirish sifatida foydalanadilar. Har bir statik tahlil atributlarini to‘plash uchun Shar PHP tilini tahlil qilish uchun maxsus ishlab chiqilgan Pixy deb nomlangan ochiq manba tahlil vositasidan foydalangan. Shundan so‘ng, Shar ushbu bashorat modelida biroz oldinga siljish uchun ma’lumotlarga bog’liqlik o‘rniga nazoratga bog’liqlik ma’lumotlaridan foydalanadigan boshqa maqola yozdi. Shunga qaramay, yorliqli zaiflik ma’lumotlarining yetishmasligi bilan shug’ullanish uchun ushbu bashorat modeliga yarim nazoratli klassifikatorni kiritish orqali tasniflash usuli bo‘yicha ko‘proq harakatlar qilindi.



1-rasm. Shar taklif qilgan dastur bayonoti darajasidagi zaiflikni bashorat qilish tizimi

Label dataset. Veb ilova har xil turdagi zaifliklarga duch keladi, lekin umuman olganda, ularni sabablariga ko‘ra ikkita asosiy turga bo‘lish mumkin, jumladan loyihalashdagi nuqson va amalga oshirishdagi xato. Shubhasiz, loyihalashdagi ko‘pgina kamchiliklarni faqat individual manba kodli fayllarni tahlil qilish orqali aniqlash qiyin. Ushbu tadqiqotda faqat zaifliklarni keltirib chiqaradigan dastur xatolari ko‘rib chiqiladi. Mashinali o‘qitish algoritmlaridan foydalangan holda ushbu ma’lumotlar namunasi bo‘yicha tasniflashda biz veb ilovaning zaifliklarini muayyan zaiflik turlariga ajratmaymiz. Bu shuni anglatadiki, ushbu tadqiqotdagi ma’lumotlar to‘plami faqat ikkilikdir, boshqacha qilib aytganda, fayl faqat zaif yoki normal bo‘lishi mumkin. Yig’ilgan fayllarni etiketlash uchun bizda ikkita manba mavjud. Ulardan biri uchinchi tomon xavfsizlik test kompaniyasi tomonidan taqdim etilgan zaiflik hisobotlari. Yana biri avtomatik xavfsizlikni tekshirish vositalari tomonidan yaratilgan hisobotlar. Ushbu ikki turdagi resurslar o‘rtasidagi asosiy farq shundaki, sinov kompaniyasining manba kodlari fayllariga kirish imkoni yo‘q, shuning uchun ular faqat qaysi veb sayt zaif ekanligini ko‘rsatishlari mumkin. Xavfsizlikni tekshirish vositalari barcha ma’lumotlar fayllariga to‘liq kirish imkoniga ega bo‘lsada, bu vositalarning aniqlash natijalari manba kodi fayliga to‘g’ri keladi.

Normallashtirish. Ushbu tadqiqotda ma’lumotlar namunalari turli ishlab chiqish paketlari va platformalaridan to‘planadi. Tasniflagichlarning ishlashini yaxshilash uchun ma’lumotlar to‘plami qiymatlar bilan mustahkam bo‘lishi kerak. Turli paketlardan yig’ilgan qiymatlar mos keladimi yoki yo‘qligini ko‘rish uchun har bir xususiyatning taqsimlanishini tasavvur qilish kerak. Xususiyat davomiyligi va kod qatorining taqsimlanishi (sloc) misol sifatida 2.8-rasmda ko‘rsatilgan. Davomiylik ko‘rsatkichi faylning mavjud bo‘lgan kunlarini ifodalaydi va qiymatga bir nechta omillar ta’sir qilishi mumkin, masalan, platforma qachon yaratilgan, ramka versiyasi qanchalik tez-tez yangilangan. Natijada, davomiylikning taqsimlanishi ishlab chiqish paketidan paketga farq qiladi. 3-paketning davomiyligi 0 dan 350 gacha, lekin 1-paketning davomiyligi 0 dan 2000 gacha taqsimlanadi. Sloc metrikasining taqsimlanishi davomiylikka o‘xshaydi, chunki metrik qiymatlarning taqsimlanishi bu

paketlar o‘rtasida ancha farq qiladi. Foydalangan turli paketlardagi nomuvofiqlik tufayli, ushbu to‘plangan ma’lumotlardan namunalarni o‘qitishdan oldin normallashtirish jarayoni zarur. Ushbu tadqiqotda quyidagi tenglama bo‘yicha x ni y ga aylantiradigan "min-max" normalizatsiya strategiyasi qo‘llaniladi:

(6)

$$y = \frac{x - \min}{\max - \min}$$

Ko‘rinishidan, normalizatsiya jarayonidan keyin ma’lumotlar to‘plami turli paketlar o‘rtasida mos keladi va qiymat 0 dan 1 gacha asl qiymat taqsimotiga ta’sir qilmasdan sodir bo‘ladi.

Xususiyat qiymati farqi. Bu yerda farqni ko‘rish uchun normalizatsiya protsedurasidan oldin ma’lumotlar to‘plamidan foydalaniladi, asosiy muammo shundaki, xususiyat qiymatlari bir-biridan juda farq qiladi, ya’ni farq qiymati bir raqamdan minglabgacha bo‘lishi mumkin. Xuddi shu rasmdagi turli ko‘rsatkichlarning farq qiymatini ko‘rsatish uchun quyidagi tenglamada aniqlanadi, bu yerda Average barcha metrik qiymatlarning o‘rtacha soni.

(7)

$$\text{Diff} = \frac{\text{Avg(Normal)} - \text{Avg(Zaif)}}{\text{Average}}$$

Har bir ko‘rsatkichning farq qiymatiga qaysi fayllar oddiy ma’lumotlar namunalari va zaif namunalar alohida kiritilganligi ta’sir qilishi mumkin. Ikki omil har bir namunaga qanday fayllar kiritilganligini aniqlashi mumkin, jumladan, turli platformalarga tegishli manba kodlari fayllari va fayllarini belgilash uchun turli zaifliklarni aniqlash vositalari tomonidan yaratilgan hisobotlardan foydalanish. Farqni ko‘rishda nazorat o‘zgarishi usulidan foydalaniladi.

Keyinchalik o‘qitishga arziydigan xususiyat, ta’sir qiluvchi omillar qanday o‘zgarishidan qat’iy nazar, har doim ijobiy (salbiy) differentsial qiymatga ega. Tez-tez o‘zgartiriladigan kunlar (num_dailychurn) va ikkita majburiyat o‘rtasidagi vaqt (commit_du) turli xil aniqlash vositalaridan foydalanish natijalari va turli platformalardagi fayllardan foydalanish orqali katta ta’sir ko‘rsatishi mumkin. Xulosa

qilib aytganda, ushbu ikki xususiyatga asoslangan bashorat natijalari ishonchsiz bo‘lishi mumkin

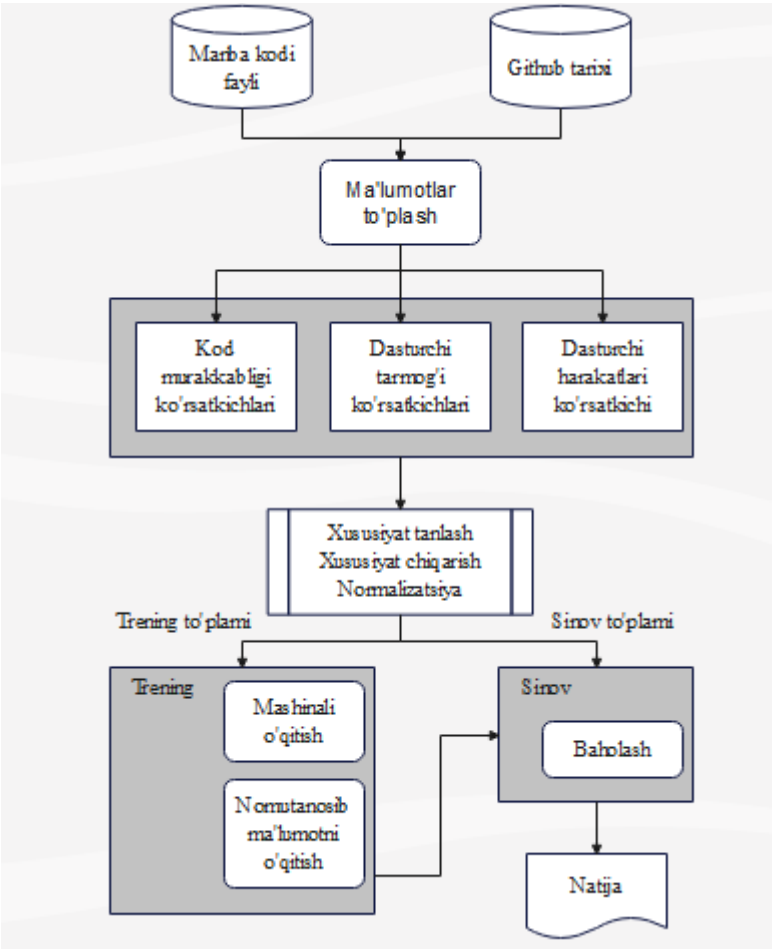
Butun jarayon quyidagi tartiblarga amal qiladi:

- Principle Component Analysis (PCA) orqali xususiyatni ajratib olish
- Trening va testlash ma’lumotlar to‘plamini yaratish
- Trening ma’lumotlari bo‘yicha ma’lumotlarni qayta namunalash
- Klassifikator parametrlarini hisoblash
- Klassifikatorlarni baholash.

Nomutanosib o‘qitish mezonlari. Mashinali o‘qitish bo‘yicha klassik tadqiqotlar bashorat qilishning aniqligini oshirishga qaratilgan. Muvozanatsiz o‘qitish uchun aniqlik turli tasniflagichlarning ishlashini solishtirish uchun yagona baholash mezonlari emas. Haqiqiy ijobiy ko‘rsatkich (TPR) va noto‘g’ri ijobiy ko‘rsatkich (FPR) ko‘pincha turli tasniflagichlarning ishlashini baholash uchun ishlatiladi. G-o‘rtacha va AUC qiymati odatda nomutanosib ma’lumotlarda bashorat qiluvchilarni yanada kengroq baholash uchun ikki sinfni muvozanatlash bo‘yicha bashoratchining ish faoliyatini o‘lchash uchun ishlatiladi. Yaxshi ishlashga ega bo‘lgan klassifikator bir vaqtning o‘zida ijobiy va salbiy sinflar uchun yuqori aniqlikka erishishi kerak, shuning uchun yaxshi tasniflagich yuqori G-o‘rtacha qiymatiga ega bo‘lishi kerak. AUC qiymati ROC egri chizig’i ostidagi maydonni o‘lchaydi va ROC egri chizig’i FPR va TPR o‘rtasidagi nisbiy almashinuvni ko‘rsatadi. ROC egri chizig’i barcha mumkin bo‘lgan qaror chegaralari bo‘ylab tasniflagichning ishlashini ko‘rsatadi. Yaxshiroq tasniflagich yuqori AUC qiymatiga ega bo‘lishi kerak. Ushbu ikki mezonga qaramasdan, balans ko‘rsatkichlari tasniflagichlarning ishlashini baholash uchun ham qo‘llaniladi. Ma’lumki, ROC egri chizig’i uchun ideal nuqta FPR = 0 va TPR = 1 bo‘lgan nuqtadir va muvozanat haqiqiy (FPR, TPR) nuqtadan (0, 1) Evklid masofasini o‘lchaydi. Balans qanchalik yuqori bo‘lsa, tasniflagich shunchalik yaxshi bo‘ladi.

(8)

$$\mathrm{G-mean} = \sqrt{\mathrm{recall}(1-\mathrm{FPR})}; \mathrm{balance} = 1 - \frac{\sqrt{(0-\mathrm{FPR})^2 + (1-\mathrm{recall})^2}}{\sqrt{2}}$$



2-rasm. Mashinali o‘qitish sxemasidan foydalangan holda zaiflikni aniqlash modeli

Parametrlarni sozlash. Zaifliklarni bashorat qilish uchun turli tasniflagichlardan foydalanish samaradorligini taqqoslashdan oldin, ta’minlash kerak. Natijani yaxshilashga yordam berish uchun nomutanosiblik muammosini hal qilish uchun ko‘proq namuna olish usulini va haddan tashqari mos kelmaslik uchun PCA usulini kiritaman. Ushbu ikki usul uchun parametrni sozlash jarayoni quyida ko‘rsatilgan.

Ko‘proq namuna olish usuli. SMOTE usuli muvozanatsiz ta’lim sohasida keng qo‘llaniladi. Ushbu maxsus haddan tashqari namuna olish usuli sintetik yangi ma’lumotlarni yaratish orqali mavjud bo‘lgan ma’lumotlar namunalarini taqsimlashni simulyatsiya qilish uchun mo‘ljallangan. Oldingi tadqiqotlarga ko‘ra, smote_ratio, bu qayta namuna olinadigan ma’lumotlar foizini nazorat qila oladigan parametr. SMOTE usuli uchun qaysi nisbat qiymati bashorat qilish modelining ish faoliyatini yaxshilashi mumkinligini aniqlash uchun. Aniq tarzda, ushbu tadqiqot uchun 5 marta o‘zaro

tekshirish usulini qo‘llash kerak. Tasodifiy vaziyatdan qochish uchun har bir yurish besh marta takrorlanadi. Hammasi bo‘lib, klassifikator bir xil ortiqcha tanlash nisbatidan foydalangan holda 5*5 natijaga ega bo‘ladi, ya’ni 2.13-rasmdagi har bir nuqta 25 ta natijaning o‘rtacha soni. Raqamlardagi tafsilotlar SMOTE nisbati qiymatining oshishi bilan birga G-o‘rtacha va muvozanatning oshishini ko‘rsatishi mumkin. Ayni paytda AUC qiymati turli SMOTE nisbati yordamida o‘zgarib turadi. Shunday qilib, SMOTE nisbati bo‘yicha qaror eng yuqori AUC qiymatini topishga bog’liq. Xulosa qilib aytganda, Naive Bayes uchun nisbat sifatida 0,4 tanlanadi; Random forest uchun 0,4 ham eng mos tanlovdir; Decision Tree uchun 0,5 ko‘proq mos keladi; Logistik diskriminant tahlili uchun 0,5 ham maqbul qiymatdir.

III. Xulosa.

Ushbu maqolada mashinali o‘qitishning ilg’or algoritmlariga asoslangan holda veb zaifliklarni oldini olish va aniqlashga qaratilgan model ishlash strukturasi ko‘rib chiqildi. U ariza maydonlari yoki URL parametrlaridan foydalangan holda dastur zaifliklarining xususiyatlarini tasniflaydi va xususiyatlarni aniqlash natijalariga ko‘ra zaifliklarni topadi va aniqlaydi. Shu bilan birga, qo‘lda aniqlashning zerikarli va qoldirilishiga yo‘l qo‘ymaslik uchun mashinani o‘rganish printsipi bilan birlashtirib, aniqlash bosqichlarini soddalashtiradi. Tajriba natijalari shuni ko‘rsatadiki, ushbu usulga asoslangan tarmoq ilovalari xavfsizligi zaifligini aniqlash dasturi yuqoridagi turdagi xavfsizlik zaifliklarini aniqlay oladi. Ammo veb-ilovalarda hali ham ba’zi xavfsizlik zaifliklari mavjud bo‘lib, ularni to‘liq aniqlash qiyin. Shuning uchun zaifliklarni aniqlash usullarini qo‘llash haqiqiy tarmoq xavfsizligi tamoyillari bilan birlashtirilishi kerak.

IV. Foydalanilgan adabiyotlar ro‘yxati:

1. Tan B , Elnaggar R , Fung J M , et al. (2020) Towards Hardware-Based IP Vulnerability Detection and Post-Deployment Patching in Systems-on-Chip[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, PP(99):1-1.

2. Yi M , Xu X , Xu L . (2019) An Intelligent Communication Warning Vulnerability Detection Algorithm Based on IoT Technology[J]. IEEE Access, 7(99):164803-164814.

3. F. Mateo Tudela, J.-R. Bermejo Higuera, J. Bermejo Higuera, J.-A. Sicilia Montalvo, and M. I. Argyros, “On combining static, dynamic and interactive analysis security testing tools to improve owasp top ten security vulnerability detection in web applications,” eng, Applied sciences, vol. 10, no. 24, pp. 9119–, 2020, ISSN: 2076-3417.

4. M. Gao, Z. Zhang, G. Yu, S. O. Arik, L. S. Davis, and T. Pfister, Consistency-based semi-supervised active learning: Towards minimizing labeling cost, 2020.

5. lBhupendra Singh Thakur, S. C. (June 2013). Content Sniffing Attack Detection in Client and Server side: A Survey. International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)Volume-3 Number-2 Issue-10 , 4.

6. Kaur A , Nayyar R . (2020) A Comparative Study of Static Code Analysis tools for Vulnerability Detection in C/C++ and JAVA Source Code[J]. Procedia Computer Science, 171(5):2023-2029.